Multi-dimensional mortality forecasting: from model averaging to neural networks

Luca De Mori

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Actuarial Science

Faculty of Actuarial Science and Insurance
Bayes Business School
City St George's, University of London

Supervisors:

Pietro Millossovich Rui Zhu Steven Haberman

January 2025

Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Pietro Millossovich, Rui Zhu, and Steven Haberman, for their guidance, valuable insights, and unwavering support, especially during difficult times. I would also like to thank the other members of the Faculty of Actuarial Science and Insurance for their inspiring comments and feedback. I am deeply grateful to my parents for the sacrifices they have made that allowed me to reach this point. To my siblings, for their constant motivation. To my friends, for encouraging me to pursue a PhD. And finally, to my colleagues, for the unforgettable moments shared during this journey.

Declaration

I hereby grant powers of discretion to the University Librarian of Bayes Business School, City St George's, University of London, to allow the thesis to be copied in whole or in part without further reference to the author. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgements.

Abstract

This dissertation addresses the challenge of jointly forecasting mortality rates across diverse populations and subgroups using a range of approaches, from traditional stochastic models to artificial neural networks. The first chapter introduces several model-averaging approaches to jointly forecast mortality rates for male and female populations within a single country, demonstrating the benefits of combining different models to improve forecasting accuracy and interval estimation. In the second chapter, the focus shifts to forecasting mortality rates for multiple populations, including both gender and country as input variables, through the application of multi-task neural networks. This approach allows for shared learning across different population groups while accounting for country-specific dynamics. The third chapter explores mortality rate predictions by cause of death and socio-economic class and their interaction using feedforward neural networks of both single-task and multi-task types. The performance of these neural networks is compared to that of existing stochastic models and machine learning approaches. Together, the chapters highlight the potential of joint modelling of mortality rates to achieve higher forecasting accuracy while pointing out their limitations and the necessity, especially for artificial neural networks, for further research and improvement.

Introduction

Mortality forecasting plays a critical role in diverse areas, including public health planning, pension systems, life insurance, and pension funds. Reliable predictions of mortality rates allow policymakers, insurers, and other stakeholders to anticipate demographic changes, allocate resources effectively, and manage financial risks associated with ageing populations. Until the publication of the work by Lee and Carter in 1992 (see Lee and Carter (1992)), which represents a milestone in the discipline of mortality forecasting, predictions of mortality-related metrics were based on simple deterministic methods and expert judgement (see Pollard (1987) for a review).
Since then, many models and approaches that provide predictions for mortality rates and other mortality- and longevity-related metrics have been proposed. Among these, both stochastic models (see Renshaw and Haberman (2006) and Cairns et al. (2006)) and, especially in the last decade, models based on machine learning (see Deprez et al. (2017) and Levantesi and Pizzorusso (2019)) can be found.
Alongside this classification of models and approaches based on their nature, it is also possible to classify them at the single-population level, where the focus is on a single population – the most studied case (see, for instance, the original Lee-Carter model) – or at the multi-population level, where the goal is to jointly forecast the metrics of interest for different populations. Among the multi-population models, the Augmented Common Factor model by Li and Lee (2005) is significant among the stochastic models, while Richman and Wüthrich (2021) proposed several deep learning models, i.e., a special case of machine learning. The scope of this dissertation falls within the multi-population approaches, with all three chapters focusing on the simultaneous forecasting of mortality rates in different populations.
This dissertation is organised into three main chapters, each of which is written as an independent and self-contained paper, plus a final chapter that serves as conclusion. They explore multi-population mortality forecasting from three different perspectives. In the following subsections of the Introduction, we provide a brief overview of each of them in terms of the problem addressed, methodology, and main results.

Chapter 1 - Two-population mortality forecasting: an approach based on model averaging

This chapter has been published as a paper in Risks journal, see De Mori et al. (2024).

The first chapter aims to introduce model-averaging approaches in a two-population mortality forecasting context and quantitatively compare them in terms of forecast accuracy with two-population versions of existing stochastic models, such as the Lee-Carter Model (see Lee and Carter (1992)), the Renshaw-Haberman Model (see Renshaw and Haberman (2006)), the Cairns-Blake-Dowd Model (see Cairns et al. (2006)), and the Augmented

Common Factor Model (see Li and Lee (2005)).
The idea behind these model-averaging approaches is that forecasts are obtained by averaging models' predictions, adopting weighting schemes of different complexity (see Shang (2012)). The advantage of using them is that we can avoid some potential drawbacks associated with using single mortality forecasting models (see Hinne et al. (2020) and Benchimol et al. (2016)). Indeed, the latter ones imply a sort of overconfidence, which consists of thinking that the selected model is the only correct one and will produce precise forecasts in any situation. This overconfidence fosters an all-or-nothing mentality, which could lead to large-scale forecasting errors (outliers). Selecting a single model could also imply incoherence: in the presence of new data (e.g., mortality rates of a different country), the selected model may no longer be optimal among those studied. Indeed, the model accuracy, on which optimal model selection depends, is heavily influenced by the dataset used in the selection process.
The two populations considered here are male and female within the same country. The model-averaging approaches, and the respective specific models, have been applied to ten different countries and a wide range of combinations of test and training periods. We used truncated life expectancy and the Gini index as metrics capturing the location and dispersion of the residual lifetime distribution. Our main conclusions are that a simple equally weighted approach performs just as well as more sophisticated averaging approaches, and model-averaging approaches are overall superior in terms of mean absolute forecasting error (see Figure 1) and, especially, interval forecast accuracy (see Figure 2).

Chapter 2 - Mortality forecasting via multi-task neural networks

A previous version of this chapter has been submitted as a paper to ASTIN Bulletin journal in July 2024, to which the reviewers answered asking for a major revision. The version corresponding to Chapter 2 has been resubmitted to the ASTIN Bulletin in January 2025.

In the second chapter, we broaden the scope by focusing on jointly forecasting mortality rates for multiple populations across various countries and both genders. To do this, we implemented several multi-task neural networks (NNs), i.e., a specific type of feedforward NN that works on different tasks simultaneously. The advantage of using NNs is that they simplify the model definition and free us from specifying how variables, such as age and calendar year, interact (see Richman and Wüthrich (2021)). Secondly, they allow us to easily consider the mortality experience of several populations simultaneously. Finally, the specificity of multi-task NNs consists of sharing their parameters among all the tasks considered, a sub-group of them, or only one of them (see Zhang and Yang 2021).
In this chapter, we quantitatively compare these multi-task NNs with pre-existing single-task NNs and stochastic models, considering mortality data from seventeen different countries. The comparison is based on mortality rates, life expectancy, and lifetime standard deviation forecasting errors. To make our analysis more complete, we considered three different age ranges and seven different training periods.
Our main conclusions are that the performance of multi-task NNs compared to single-task NNs and stochastic models is highly dependent on the metric, age range, and training period considered, see Figures 3 and 4. Overall, single-task NNs give the best results in terms of mortality rates forecasting error, while multi-task NNs and stochastic models have the lowest forecasting error, respectively, for life expectancy and lifetime standard

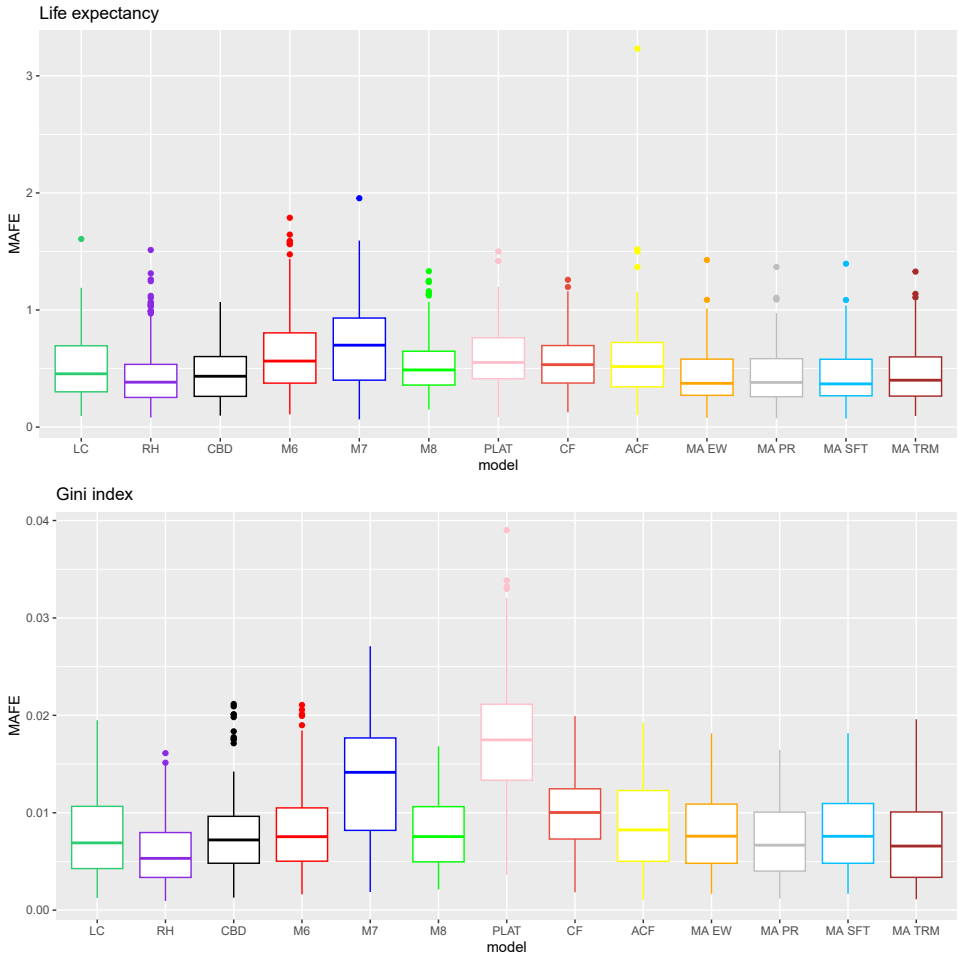


Figure 1: Mean absolute forecasting error (MAFE) for life expectancy and Gini index. Results for individual stochastic models (LC-ACF) and according weighted average (MA EW-MA TRM). 10 countries and 26 test periods considered here.

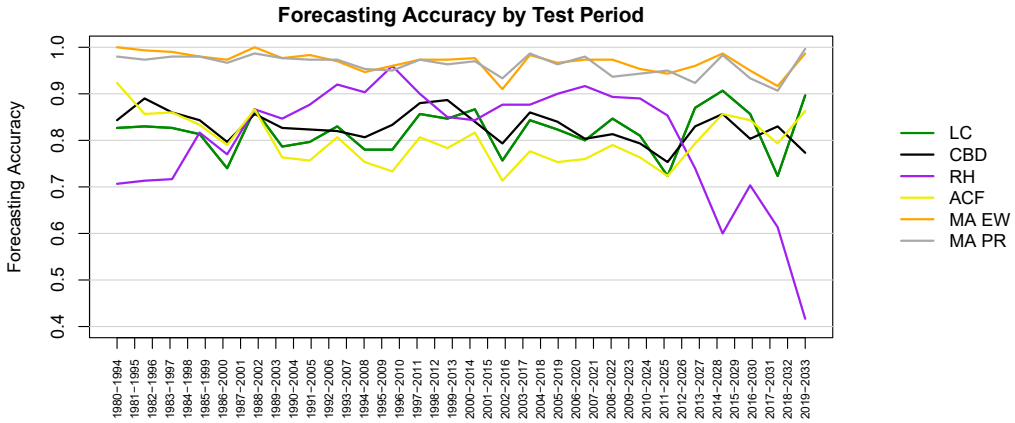
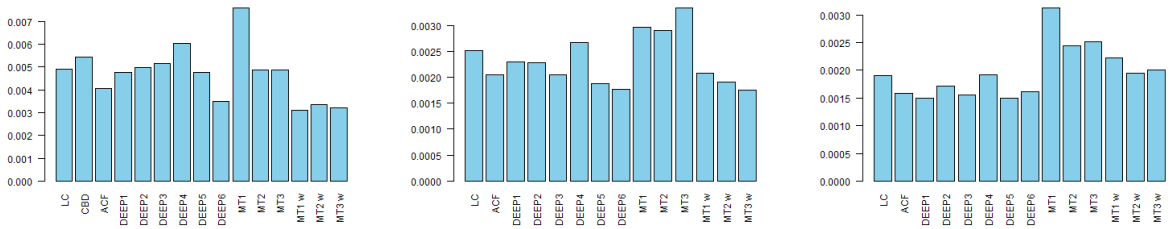
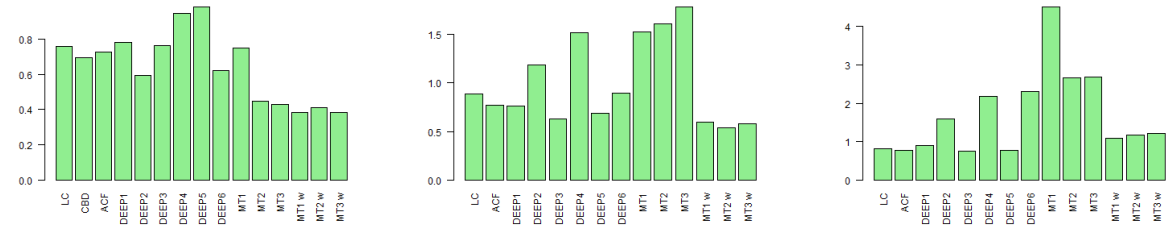


Figure 2: Interval forecasting accuracy by test period for Lee-Carter (LC), Cairns-Blake-Dowd (CBD), Renshaw-Haberman (RH), and Augmented Common Factor (ACF) models, and model averaging approaches with equal (MA EW) and proportional (MA PR) weights. Proportion of cases in which the observed life expectancy falls in the forecasting interval.

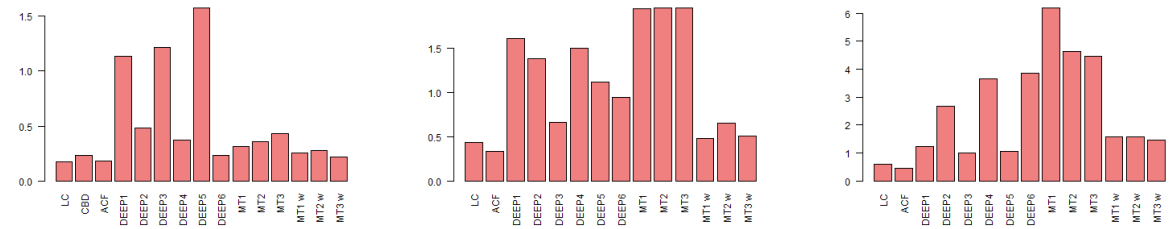
deviation. Furthermore, implementing a weighting scheme improves the performance of multi-task NNs, especially for life expectancy and lifetime standard deviation when considering wider age ranges.



(a) Mort. Rates (55-89) (b) Mort. Rates (20-89) (c) Mort. Rates (0-89)



(d) Life Exp. (55-89) (e) Life Exp. (20-89) (f) Life Exp. (0-89)



(g) Std. Dev. (55-89) (h) Std. Dev. (20-89) (i) Std. Dev. (0-89)

Figure 3: Mean absolute forecasting error (MAFE) for Mortality Rates, Life Expectancy, and Standard Deviation by approach across different age ranges. LC, CBD, and ACF are stochastic models; DEEP1-DEEP6 are single-task neural networks; MT1-MT3 and MT1 w-MT3 are multi-task neural networks respectively without and with a weighting scheme.

Chapter 3 - Forecasting mortality rates by cause of death and socio-economic class using neural networks

The third chapter shifts the focus to mortality within a single country, considering how it changes by cause of death and socio-economic class. The impact of these two variables on mortality rates is considered both individually and conjointly. More specifically, we aim to implement feedforward neural networks, both of the single-task and multi-task types, for forecasting mortality rates by single causes of death (rather than the overall ones) and by socio-economic class. Additionally, the chapter briefly introduces other methodologies

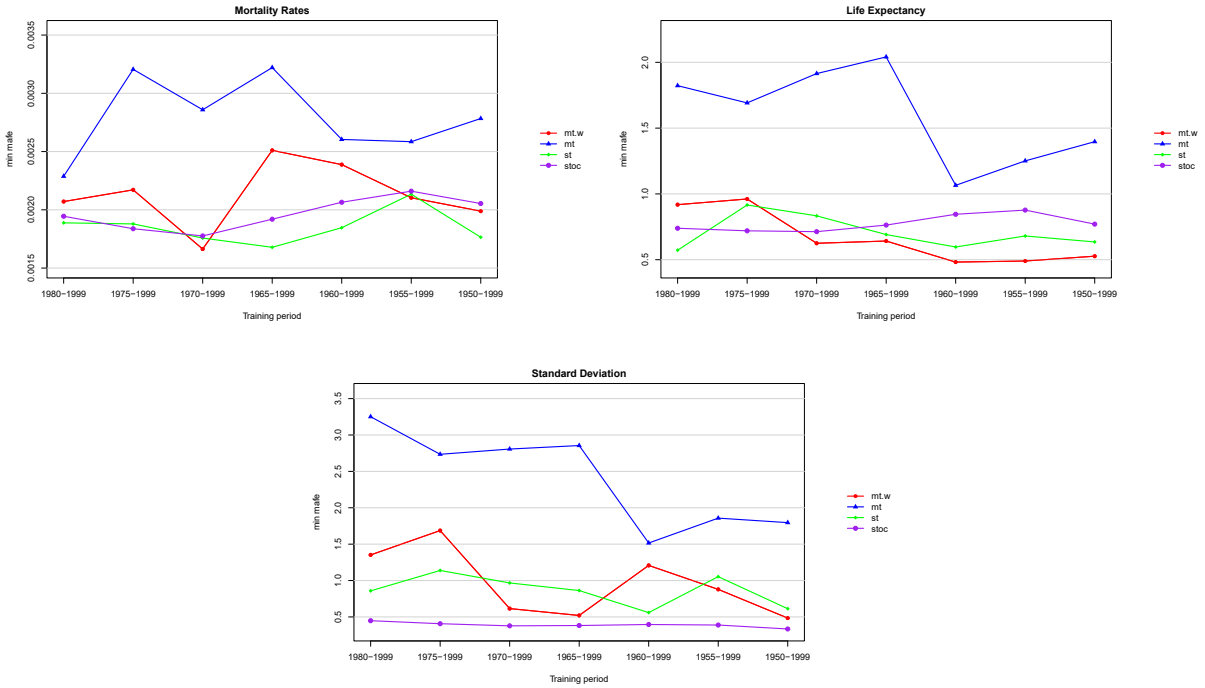


Figure 4: Minimum mean absolute forecast error (MAFE) among stochastic models, single-task NNs, and multi-task NNs (both with and without a weighting scheme), by training period and metric considered. Test period: 2000-2019.

in the literature that have been utilised for the same purpose: the Lee-Carter model and Penalised Tensor Decomposition (see Zhang et al. (2023)). Finally, we test both neural networks and alternative methodologies on an England dataset to evaluate their forecasting performance in terms of mean absolute error, mean squared error, and mean absolute percentage error.

There are two main conclusions that emerged from the quantitative analysis here. The first is that neural networks can achieve similar results in terms of MSE and especially MAE compared with the Lee-Carter model, i.e., the leading model in the literature for out-of-sample performance by cause of death. The second finding is that neural networks have poorer performance in terms of MAPE compared to the Lee-Carter model and Penalised Tensor Decomposition. Subsequently, we implemented a weighting scheme in the training of the neural networks to reduce the magnitude of the error in the lower age classes, where the MAPE was particularly high. We conclude that this implementation notably improves the results in terms of MAPE while not significantly changing the results in terms of MSE and MAE, see Figure 5.

Figure 5: Comparison of performance between neural networks with and without a weighting scheme by metric and case. The metrics considered are mean absolute error (MAE), mean absolute percentage error (MAPE), and mean squared error (MSE). The cases considered are 4 causes of death, 6 causes of death, 5 socio-economic classes (IMD), and 6 causes of death and 5 socio-economic classes jointly.

Chapter 1

Two-population mortality forecasting: an approach based on model averaging

1.1 Introduction

In recent decades, as a consequence of life expectancy improvements and social and behavioural changes that have taken place in various countries, pension funds, annuities, and other insurance products that provide retirement income have become increasingly important. However, since these products are subjected to longevity risk, which refers to the systematic trend of mortality rates decreasing over time, it has become necessary to find effective models for forecasting mortality rates. Notably, several models have been developed to address this issue, including the Lee-Carter model (1992), its extension the Renshaw-Haberman model (2006), the Cairns-Blake-Dowd model (2006), and its extensions - the M6, M7, and M8 models (2009) - and the Plat model (2009). In all these models, the mortality rates depend on two types of quantities: fixed parameters that represent the effect of age on mortality; and stochastic factors that represent the effect of cohort year and calendar year. All these quantities must be estimated using statistical techniques based on past data. Finally, the mortality rates are forecasted extrapolating the stochastic factors on the more recent period. Originally, these models had been designed to forecast the mortality rates of single populations. Later, they were implemented in a multi-population framework by Li & Lee (2005). This class of models has become increasingly popular because it allows researchers to work simultaneously with different populations that are to some extent related (e.g., males and females in the same country or region) obtaining coherent forecasts, see Dowd et al. (2011), Li (2013), Yang et al. (2016), Enchev et al. (2017), and Shang et al. (2022). This means that these models are able to respect limitations and constraints that we set such as the biological ones like the sex gap between the life expectancy of females and males. As an instance of practical application, we can refer to the longevity risk transfer products where it is necessary to quantify the basis risk that is the systematic difference between the population mortality and the pension fund mortality, see Villegas et al. (2017).

As researchers continue to make progress in developing single-population and multi-population mortality forecasting models, there has also been a growing interest in model averaging approaches in the last years, see Shang et al. (2011), Shang (2012), and Benchimol et al. (2016). The idea behind them is that forecasts are obtained by av-

eraging models' predictions using various weighting schemes. By adopting a model averaging approach, we can avoid some potential drawbacks derived by using single or multi-population mortality forecasting models, see Hinne et al. (2020), and Benchimol et al. (2016). They imply a sort of over-confidence, that consists of thinking that the selected model is the only one correct and will produce precise forecasts in any situation. They imply an all-or-nothing mentality. They could lead to big-scale forecasting errors (outliers). They could imply incoherence: in the presence of new data (mortality rates of a different country, for instance), the selected model may no longer be optimal among those studied. Indeed, the model accuracy, on which optimal model selection depends, is heavily influenced by the dataset used in the selection process.
The novelty of this paper lies in introducing model averaging approaches within a two-population mortality forecasting context, whereas existing studies in the literature mainly focus on forecasting mortality at the single-population level. Furthermore, it quantitatively evaluates these approaches, highlighting their differences and comparing them with traditional single two-population models. We used truncated life expectancy and the Gini index as metrics capturing the location and dispersion of the residual lifetime distribution. Our main conclusions are that a simple, equally weighted approach performs just as well as more sophisticated averaging approaches, and that model averaging approaches are, overall, superior in terms of mean absolute forecasting error and interval forecast accuracy to most common two-population models, when considering a range of combinations of test and training periods and countries.
The remainder of this paper is organized as follows. In Section 2, we revisit some existing stochastic multi-population models in literature with a particular focus on the two-population case. In Section 3, model-averaging approaches applied to two-population mortality forecasting are introduced from a theoretical point of view. In Section 4, we list the data sets that we use in our practical examples. In Section 5, the procedure implemented to obtain the quantitative results is described step-by-step. In Section 6, the results are presented and discussed. Finally, in Section 7, we summarize the most important findings of the previous sections and provide some future outlooks.

1.2 Two-population mortality models

Let us consider two populations, denoted by $p = m, f$, that represent males and females of one specific country, and assume that for both of them, the distribution of deaths between successive integer ages is uniform. The number of deaths $D_{x,t}^{(p)}$ in population p, year t, and age x, conditionally on the central death rate $m_{x,t}^{(p)}$, is assumed to follow a Poisson distribution:

$$D_{x,t}^{(p)} \sim Po(N_{x,t}^{(p)} m_{x,t}^{(p)}), \tag{1.1}$$

where $N_{x,t}^{(p)}$ is the central exposure to risk. In order to forecast central death rates several models have been proposed in the last three decades following the original idea of Lee and Carter (1992). In all these models, the natural logarithm of mortality rates $m_{x,t}^{(p)}$, or of probability of death $q_{x,t}^{(p)}$,[1] is expressed as function of two different types of quantities: time-dependent stochastic factors and age-dependent parameters. A relevant selection of these models is considered in this paper in their two-population form and summarized in

[1]Probabilities of death q_x can be calculated from the corresponding mortality rates m_x by using the relation $q_x = m_x/(1 + \frac{1}{2}m_x)$, and vice versa, $m_x = q_x/(1 - \frac{1}{2}q_x)$.

Table 1.1.[2] More precisely, in any given population, the mortality rates of both females ($m^f_{x,t}$) and males ($m^m_{x,t}$) are specified by one of the equations in 1.-9. in Table 1. These models can be broadly classified as follows: models where the age is treated as categorical (1, 2, 8 and 9), as a quantitative variable (3-6) or hybrid (7); models which consider a cohort effect (2, 4 to 7) or not; models with a stochastic time factor common to both females and males populations (8 and 9) or models where the dependence between sexes only stems from the correlation in stochastic time factors; models with one (1, 2 and 8), two (3, 4, 6, 7 and 9) or three stochastic time factors (5).

Table 1.1: Summary of multi-population models used. Here $\kappa^{(i,p)}_t$, $i = 1, 2, 3$, κ_t, and $\kappa^{(p)}_t$ are time-varying stochastic factors, $\gamma^{(p)}_{t-x}$ are cohort-related stochastic factors, $\beta^{(i,p)}_x$, $i = 1, 2, 3$, and $\beta^{(2)}_x$ are age-specific parameters, $\bar{x} = \frac{1}{m+1}\sum_{i=0}^{m} x_i$ is the mean age over the population age range, $\hat{\sigma}^2_x = \frac{1}{m+1}\sum_{i=0}^{m}(x_i - \bar{x})^2$ is the age variance, and finally $x^{(p)}_c$ is an arbitrary fixed age.

Model	$\ln(m^{(p)}_{x,t})$
1. Lee-Carter model (LC)	$\beta^{(1,p)}_x + \beta^{(2,p)}_x \kappa^{(2,p)}_t$
2. Renshaw-Haberman model (RH)	$\beta^{(1,p)}_x + \beta^{(2,p)}_x \kappa^{(2,p)}_t + \gamma^{(p)}_{t-x}$
3. Cairns-Blake-Dowd model (CBD)	$\kappa^{(1,p)}_t + \kappa^{(2,p)}_t (x - \bar{x})$
4. CBD Model with a cohort Effect (M6)	$\kappa^{(1,p)}_t + \kappa^{(2,p)}_t (x - \bar{x}) + \gamma^{(p)}_{t-x}$
5. CBD Model with quadratic and cohort effects (M7)	$\kappa^{(1,p)}_t + \kappa^{(2,p)}_t (x - \bar{x}) + \kappa^{(3,p)}_t ((x - \bar{x})^2 - \hat{\sigma}^2_x) + \gamma^{(p)}_{t-x}$
6. CBD Model with an age-dependent cohort effect (M8)	$\kappa^{(1,p)}_t + \kappa^{(2,p)}_t (x - \bar{x}) + \gamma^{(p)}_{t-x}(x^{(p)}_c - x)$
7. Plat Model (PLAT)	$\beta^{(1,p)}_x + \kappa^{(1,p)}_t + \kappa^{(2,p)}_t (x - \bar{x}) + \gamma^{(p)}_{t-x}$
8. Common Factor Model (CF)	$\beta^{(1,p)}_x + \beta^{(2)}_x \kappa_t$
9. Augmented Common Factor Model (ACF)	$\beta^{(1,p)}_x + \beta^{(2)}_x \kappa_t + \beta^{(2,p)}_x \kappa^{(2,p)}_t$

[2]For consistency, we use the Poisson distribution assumption coupled with the log-link function and mortality rates for models such as M6, M7, and M8 which usually are presented under a binomial assumption coupled with the logit-link function and probabilities of death.

1.2.1 Model estimation

The parameters of the models in Table 1.1 are usually estimated by maximizing the joint Poisson log-likelihood:

$$\ell = \sum_p \sum_x \sum_t \left\{ d_{x,t}^{(p)} \ln(N_{x,t}^{(p)} m_{x,t}^{(p)}) - N_{x,t}^{(p)} m_{x,t}^{(p)} - \ln(d_{x,t}^{(p)}!) \right\} \tag{1.2}$$

where $d_{x,t}^{(p)}$ are the observed deaths in population p, year t, and age x. The mortality rates $m_{x,t}^{(p)}$ for each model can be obtained from the corresponding equations in Table 1.1. The optimization is performed using numerical algorithms. Note that for models 1-7, the log-likelihood for each population can be maximized separately.

1.2.2 Stochastic factors assumptions

From Table 1.1, it can be seen that the models considered depend on a number of stochastic factors. More precisely, each model contains a combination of (one or more of) the following terms: population specific time indices $\kappa_t^{(i,p)}$, common time index κ_t and population specific cohort effects $\gamma_t^{(p)}$. Inspired by Li et al (2015), for the time indices $\kappa_t^{(i,p)}$ (models 1-7, 9) we consider a combination of a random walk with drift and first-order autoregression AR(1). The rationale of this choice is that there is a stable relation between the period indices of males and females.

- $\kappa_t^{(i,m)} = \mu^{(i,m)} + \kappa_{t-1}^{(i,m)} + Z_t^{(i,m)}, \quad i = 1, 2, 3$
- $\kappa_t^{(i,f)} = \kappa_t^{(i,m)} + \phi^{(i,f)}(\kappa_{t-1}^{(i,f)} - \kappa_{t-1}^{(i,m)}) + Z_t^{(i,f)}, \quad i = 1, 2, 3$

where $\mu^{(i,m)}$ are the drift parameters, $\phi^{(i,f)}$ are the autoregressive parameters, and $(Z_t^{(i,p)})_{p=m,f}$ are normal iid innovations.

For the time index κ_t (models 8-9), we consider a random walk with drift:

- $\kappa_t = \mu + \kappa_{t-1} + Z_t,$

where μ is the drift parameter, and Z_t are normal iid innovations.

Finally, for the cohort terms $\gamma_{t-x}^{(p)}$ (models 2, 4-7), we consider a combination of ARIMA(1,1,0), see Villegas et al. (2017), and Dowd et al. (2010), and first-order autoregression AR(1), see Li et al. (2015). Again, the rationale is that there is a stable relation between the cohort effects of males and females.

- $\gamma_u^{(m)} = (1 + \phi^{(m)})\gamma_{u-1}^{(m)} - \phi^{(m)}\gamma_{u-2}^{(m)} + Y_u^{(m)}$
- $\gamma_u^{(f)} = \gamma_u^{(m)} + \phi^{(f)}(\gamma_{u-1}^{(f)} - \gamma_{u-1}^{(m)}) + Y_u^{(f)}$

where $\phi^{(m)}$ and $\phi^{(f)}$ are the autoregressive parameters of the process, while $(Y_u^{(p)})_{p=m,f}$ are normal iid innovations.

In models 1-7, the dependence between female and male mortality is derived from the autoregressive components. In model 8, the dependence is given by the shared time index κ_t between female and male populations. In model 9, the dependence is derived by both the shared time index and the autoregressive component.

1.3 Model averaging approaches

Suppose we have historical data on mortality for the period $[t_0, t_s]$, and we are interested in forecasting some mortality metric on the period $[t_{s+1}, t_n]$. The purpose of model averaging approaches is to obtain forecasts of a given metric

$$U_t^{(p)} = f((m_{x,t}^{(p)})_{x=0,\dots,\omega-1}) \tag{1.3}$$

that can be expressed as a function of mortality rates, where ω is the ultimate age, as an average of the forecasted metrics obtained using L different models. Notice that in this paper, we consider individual models listed in Table 1, and $L = 9$. As an example of the metric $U_t^{(p)}$, we can think to the j-years survival probability ${}_jp_{x,t}^{(p)} = \exp\{-(m_{x,t}^{(p)} + m_{x+1,t}^{(p)} + \cdots + m_{x+j-1,t}^{(p)})\}$.
Let the metric of interest for the population p, in year t, and model l be

$$\hat{U}_t^{(p,l)} = f((\hat{m}_{x,t}^{(p,l)})_{x=0,\dots,\omega-1}), \qquad l = 1, \dots, L \tag{1.4}$$

where $\hat{m}_{x,t}^{(p,l)}$ is the forecasted mortality rate at age x, for the population p, in year t, obtained using model l. Following Fletcher (2018) and Shang (2012), the averaged metric is calculated as

$$\hat{U}_t^{(p,average)} = \lambda_1 \hat{U}_t^{(p,1)} + \cdots + \lambda_L \hat{U}_t^{(p,L)} \tag{1.5}$$

where $\lambda_1, \dots, \lambda_L$ are non-negative weights calculated based on the model averaging approach considered dependently on the performance of the models in the validation period. In this paper, we consider the following four model averaging approaches:

- Equal weights (EW):

$$\lambda_l^{EW} = \frac{1}{L}, \qquad l = 1, \dots, L. \tag{1.6}$$

  It is the most simple, as all models are assigned the same weight, see Shang (2012). There is no penalisation or reward depending on the performance in the validation period.

- Proportional weights (PW):

$$\lambda_l^{PR} = \frac{\frac{1}{g_l}}{\sum_{k=1}^{L} \frac{1}{g_k}}, \qquad l = 1, \dots, L \tag{1.7}$$

  where $g_l = g(\hat{U}_t^{(p,l)}, U_t^{(p,l)}; p = m, f; t = t_s - h + 1, \dots, t_s)$ is a strictly positive performance measure representing the performance of the model l in the validation period $[t_s - h + 1, t_s]$, see Shang (2012). In this way, models that have poor performance in the validation period are penalised with smaller weights.

- Weights based on the SoftMax function (SM):

$$\lambda_l^{SM} = \frac{\exp\{-g_l\}}{\sum_{k=1}^{L} \exp\{-g_k\}}, \qquad l = 1, \dots, L. \tag{1.8}$$

  The concept is similar to the proportional weights model averaging approach, but here we penalise less the models with poor performances in the validation period and reward less the models with good performances. See also Benchimol et al. (2016) for a similar formulation.

- Weights based on trimming (TR):

$$\lambda_l^{TR} = \begin{cases} \frac{1}{\hat{L}}, & \text{if } l \text{ is among the } \hat{L} \text{ best models in the validation period,} \\ 0, & \text{otherwise,} \end{cases} \tag{1.9}$$

  where the best models are determined in terms of the measure g_l, see Samuels and Sekkel (2017), and Shang (2012). With this method, we reward only the $\hat{L}$ models that have the best performances in the validation period assigning the same weight ($\frac{1}{\hat{L}}$) for each one of them. In the following, we set $\hat{L} = 3$, representing an intermediate choice between 1 — corresponding to selecting solely the model that achieves the best performance during the validation period — and 9 — corresponding to averaging across all considered models with equal weights, which is equivalent to EW approach.

In the remainder of this paper, for the definition of measure g used to evaluate the models' performance in the validation period, we adopt the mean absolute forecasting error (MAFE)

$$\text{MAFE}_l = \frac{\sum_{p=m,f} \sum_{t=t_s-h+1}^{t_s} |\hat{U}_t^{(p,l)} - U_t^{(p,l)}|}{P \cdot h} \tag{1.10}$$

where t_{s-h+1} and t_s are the first and last years of the validation period, and P is the total number of populations considered. Notice that alternative measures could also have been considered. For example, one that also takes into account the number of parameters in the model.

Finally, regarding the metric in (1.4), we choose the residual life expectancy at age 55 truncated at age 90, which represents a location metric of mortality rates, see Dickson et al. (2019)

$$\mathring{e}_{55:\overline{35}|,t} = \sum_{j=1}^{35} {}_{j-1}p_{55,t}(1 - \frac{1}{2}q_{55+j-1,t}) \tag{1.11}$$

and the Gini index, calculated between 55 and 89, that represents the dispersion of mortality rates[3]

$$G_{55:\overline{35}|,t} = \frac{1}{2\mathring{e}_{55:\overline{35}|,t}} \Big(\sum_{x=55}^{89} \sum_{y=55}^{89} {}_{x-55|1}q_{55,t} \; {}_{y-55|1}q_{55,t} \mid x - y \mid \tag{1.12}$$

$$+ 2 \sum_{x=55}^{89} {}_{x-55|1}q_{55,t} \; {}_{35}p_{55,t} \mid x - 90 \mid \Big). \tag{1.13}$$

The Gini index is a metric that varies between the limits of 0 (perfect equality) and 1 (perfect inequality). In our case, given a population of n individuals, the Gini index is equal to zero if all n individuals die at the same age, while it converges to one if $n-1$ individuals die at age 55 and a single individual dies after age 89, as n tends to infinity. The choice of the Gini index as a metric representing the dispersion of mortality depends on the fact that, unlike other metrics such as interquartile range, variance, and standard deviation, it possesses all the desirable properties for an inequality index: population-size independence, mean or scale independence, and transfer principle, see Shkolnikov et al. (2003) for more details.

[3]In the paper published in the journal *Risks* (2024), we used an approximation given by: $G_{55:\overline{35}|,t} = \frac{1}{2\mathring{e}_{55:\overline{35}|,t}} \sum_{x=55}^{89} \sum_{y=55}^{89} {}_{x-55|1}q_{55,t} \; {}_{y-55|1}q_{55,t} \mid x-y \mid$. This approximation works well when the probability of surviving beyond age 89 is not too high.

1.4 Data

In the following, we implement the multi-population models and the model averaging approaches with historical mortality data from ten pairs of populations, namely the female and male populations of Australia, Canada, France, England & Wales, Italy, Japan, Netherlands, Spain, Sweden, and the US. The choice of these countries depends on the fact that they are developed with large populations, and their data are complete and easily obtainable in the Human Mortality Database (HMD). For these countries, we considered the age range between 55 and 89 for the mortality rates. The reason for this choice derives from the fact that most of the deaths are concentrated after the age of 55 years and that after the age of 90 years, we have less data, especially for the older cohorts, so this could lead to biased estimations of the models. Also, we note that, at ages over 90, there is evidence of age misstatements that may lead to biased estimates of mortality indices. Furthermore, this age range is the most relevant from an actuarial point of view, see Cairns et al. (2006). Finally, concerning the periods considered we have two cases: in the first one, we have 30 years rolling training period from 1950-1979 to 1975-2004 and 15 years rolling test period from 1980-1994 to 2005-2019 (last year on which the data are available at the epoch in which we are writing this paper); in the second one, we have variable length training period from 1966-2004 to 1985-2004, and 15 years fixed test period 2005-2019.

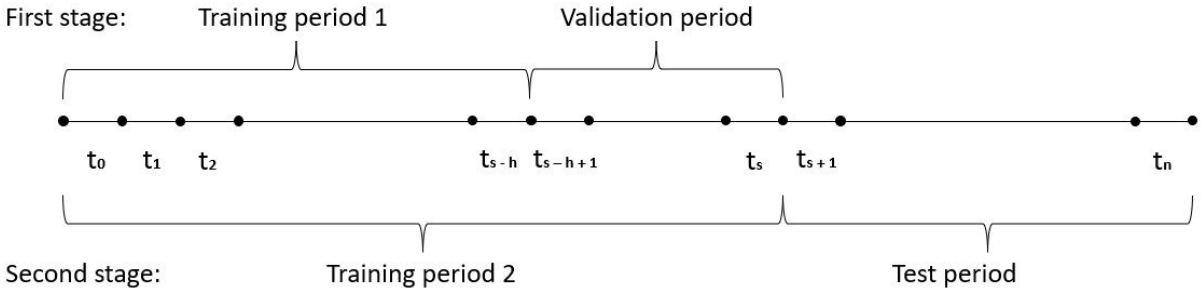


Figure 1.1: An illustration of the training, validation, and test periods to train and evaluate the models.

1.5 Implementation

1.5.1 Step by step procedure

For a generic country, training period $[t_0, t_s]$, and test period $[t_s + 1, t_n]$ (see Figure 1.1), we follow the steps listed below:

1. First stage

   1.1 We fit the two-population models (LC, RH, CBD, PLAT, M6, M7, M8, CF, and ACF) on the period $[t_0, t_s - 10]$ (training period 1) using the StMoMo package, see Villegas et al. (2018). Notice that this implies h, i.e. the length of the validation period, is set equal to 10.

   1.2 We simulate mortality rates for the period $[t_s - 9, t_s]$ (validation period) using the models fitted in 1.1.

1.3 We calculate the corresponding truncated life expectancy and Gini index for each model as functions of the mortality rates obtained in 1.2 using Formulas (1.11) and (1.12).

1.4 We repeat steps 1.2 and 1.3 for 1000 times and we obtain the forecasted truncated life expectancy and Gini index as the average of these for each model.

1.5 We calculate the MAFE as the difference between forecasted truncated life expectancy and Gini index calculated in 1.4 and the historical ones (Formula (1.10)) for each model.

1.6 We calculate the weights of each model averaging approach based on the MAFEs calculated in 1.5 using Formulas (1.6), (1.7), (1.8), and (1.9).

2. Second stage

2.1 We repeat step 1.1 using the period $[t_0, t_s]$ (training period 2) instead of $[t_0, t_s - 10]$.

2.2 We repeat steps 1.2 and 1.3 using the period $[t_s + 1, t_n]$ (test period) instead of $[t_s - 9, t_s]$.

2.3 We repeat step 2.2 for 10000 times[4] and we obtain the forecasted truncated average life expectancy and Gini index as the average of these for each model.

2.4 For each model averaging approach, we make 1 simulation from a multinomial distribution with parameters equal to 10000, 9, and the vector of the weights obtained in 1.6. The result of this simulation will be a vector with 9 elements, which sum to 10000, that represent the number of truncated life expectancy and Gini index trajectories that are considered in the model averaging approach from the 9 two-population models.

2.5 Using the results of the simulation at point 2.4 as parameters, we resample by bootstrap from the truncated life expectancy and Gini index trajectories obtained in step 2.3, so in total we will have 10000 trajectories for both metrics. Then we average them using Formula (1.5) obtaining the forecasted truncated life expectancy and Gini index for all the model averaging approaches. Similarly, we take the 5th and 95th percentile from the set of 10000 trajectories to build the 90% confidence forecasting intervals for the two metrics. See Figure 1.2 for an example of forecasted life expectancy and Gini index using the model averaging approach with equal weights.

2.6 We calculate the MAFE as the difference between the forecasted truncated life expectancy and the Gini index calculated in 2.3 and 2.5 with the observed ones. Similarly, we determine the interval forecast accuracy of the two metrics as the proportion of times in which the observed truncated life expectancy and Gini index fall within the respective confidence forecasting intervals.

1.6 Results

Following Shang (2012), in order to evaluate the performances of each model, we want to consider the goodness of both point and interval forecasts. For the first one, we consider

[4]We make more simulations than in the first stage since here we consider the interval forecast accuracy in addition to the MAFE.

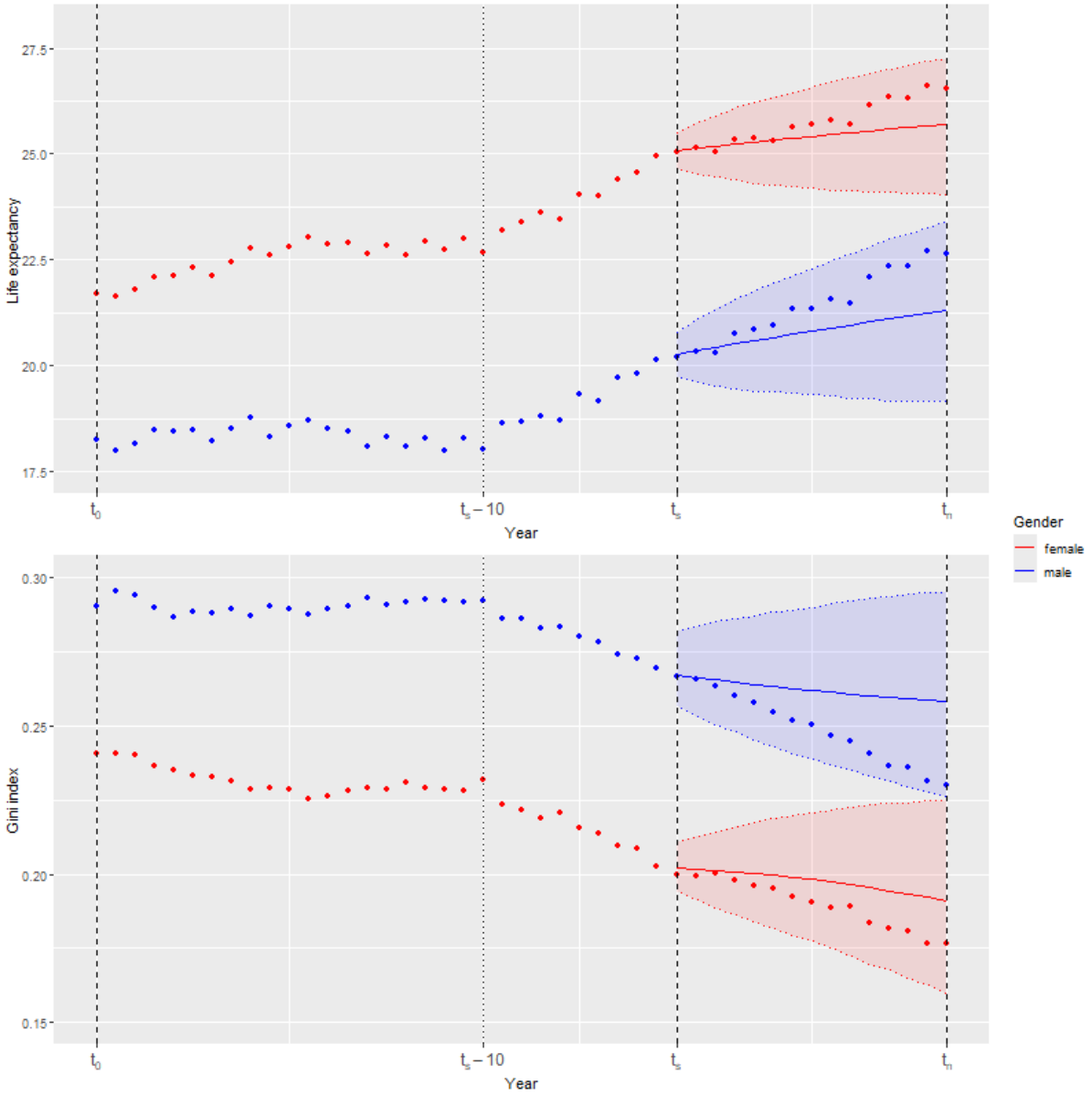


Figure 1.2: Example of forecasted truncated life expectancy and Gini index, the respective 90% prediction intervals, and the observed values of the two metrics (points). Training period: 1950-1979. Test period: 1980-1994. Country: Australia. Model averaging approach: equal weights.

the mean absolute forecasting error (MAFE), see Formula (1.10), while for the second one, the interval forecast accuracy, here defined as the proportion of cases in which the observed life expectancy or Gini index falls within the 90% confidence forecasting interval. In Figures 1.3 and 1.5, we can find boxplots summarising the mean absolute forecasting errors for life expectancy and the Gini index obtained by all the models and model averaging approaches previously mentioned, respectively in the rolling and fixed test period cases. Figure 1.4 shows, for the rolling test period case, which is the best model, i.e. the one with the lowest MAFE, by period, country, and metric. Figure 1.6 has the same content but for the fixed test period case. Finally, Tables 1.2-1.5 and 1.6-1.9 report the interval forecast accuracy of the models by period, country, and metric, respectively for rolling and fixed test period cases.

1.6.1 Rolling test period

Observing Figure 1.3, we can generally conclude that the best results, in terms of MAFE, are given by the model averaging approaches with equal and proportional weights, and based on softmax function, alongside with the RH model. They all present a median MAFE lower than 0.4 and 0.008 respectively for life expectancy and the Gini index. The RH model is the third best model (after model averaging approaches with equal weights, and with weights based on softmax function) for life expectancy, and the best model for the Gini index. The model averaging approach based on trimming has good overall performances as well. Indeed, it is overcome in terms of median MAFE only by the RH model for both life expectancy and the Gini index. Among the other models, good results are given by the CBD model for what concerns life expectancy and by the LC model for the Gini index. The worst results here are found in the M7 model for life expectancy and the PLAT model for the Gini index. Other models, such as M6, CF, and ACF, do not show remarkable results.

Figure 1.4 shows the model with the lowest MAFE by country and period. The RH model has the highest number of best performances for both life expectancy and the Gini index (34% and 35%),[5] followed by the CBD model (17% and 16%). It is interesting to notice how here the M7 model is the third best model for life expectancy, despite it being the worst one in terms of median MAFE. Focusing on the model averaging approaches, it can be noticed that they are seldom the best ones. Among them, the trimming averaging approach has the highest proportion of winning cases (7% and 17%). These results, even if they appear to contradict those in Figure 1.3, are easily explained. Indeed, model averaging approaches consistently perform well, ranking among the top five or six positions for lowest MAFE across all countries, periods, and metrics. On the contrary, it happens that individual models that are the best for specific combinations of countries and periods, such as M7 for life expectancy, are overall among the worst, as shown in Figure 1.3. Consequently, the advantage of model averaging is striking, in particular when several countries must be considered and models must be updated on different training periods. Tables 1.2, 1.3, 1.4, and 1.5 show the interval forecast accuracy results. As happened for the median MAFE analysis, the model averaging approaches based on equal and proportional weights, and on the softmax function are the best ones with an interval forecast accuracy of 97% for what concerns the life expectancy, and between 92% and 93% for

[5]These percentages have been calculated as the ratio between the number of cases in which each model is the best over the total number of cases considered (260).

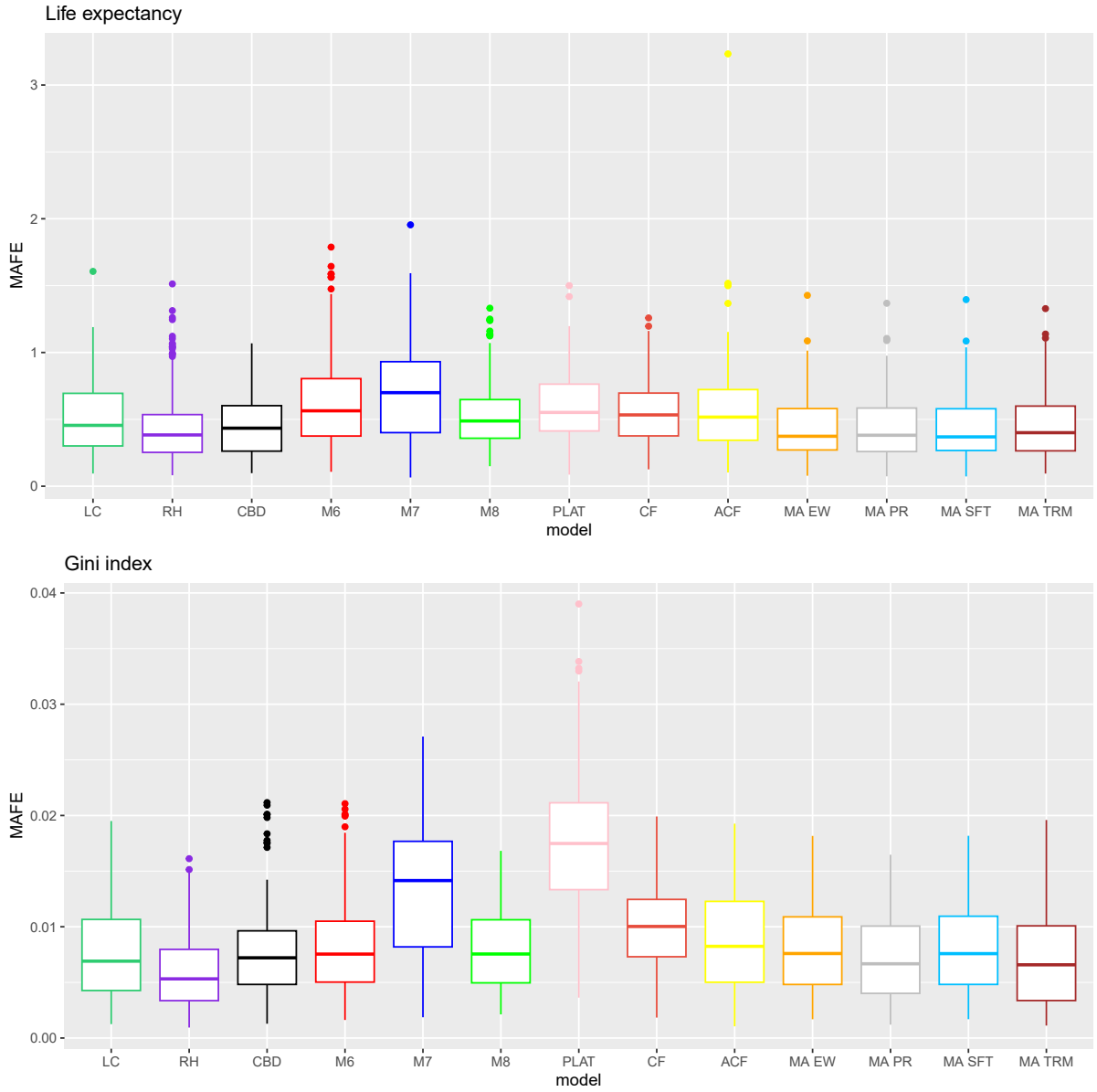


Figure 1.3: Summary of the MAFEs by model. Results for individual models and according weighted average. Rolling test period case.

what concerns the Gini index. They are followed by the model averaging approach based on trimming (93% and 94%). Among the other models, we have that the RH model has an interval forecast accuracy of 81% for the life expectancy (lower than the LC and CBD models), and of 75% for the Gini index (lower than the LC model).

In the following, we report other findings from the tables and figures mentioned above, focusing particularly on the analysis by country, period, and metric considered. In Figure 1.4, we notice how the best model changes over time due to changes in mortality trends, such as the slowing down in mortality improvements observed in several developed countries since 2010, see Djeundje et al. (2022). Similar behaviour can also be noticed in Tables 1.2 and 1.4 for the forecast interval accuracy. In this regard, the evolution of the RH model's performances is enlightening. Indeed, it goes from being the best single model for life expectancy in many test periods considered, to becoming the worst one in

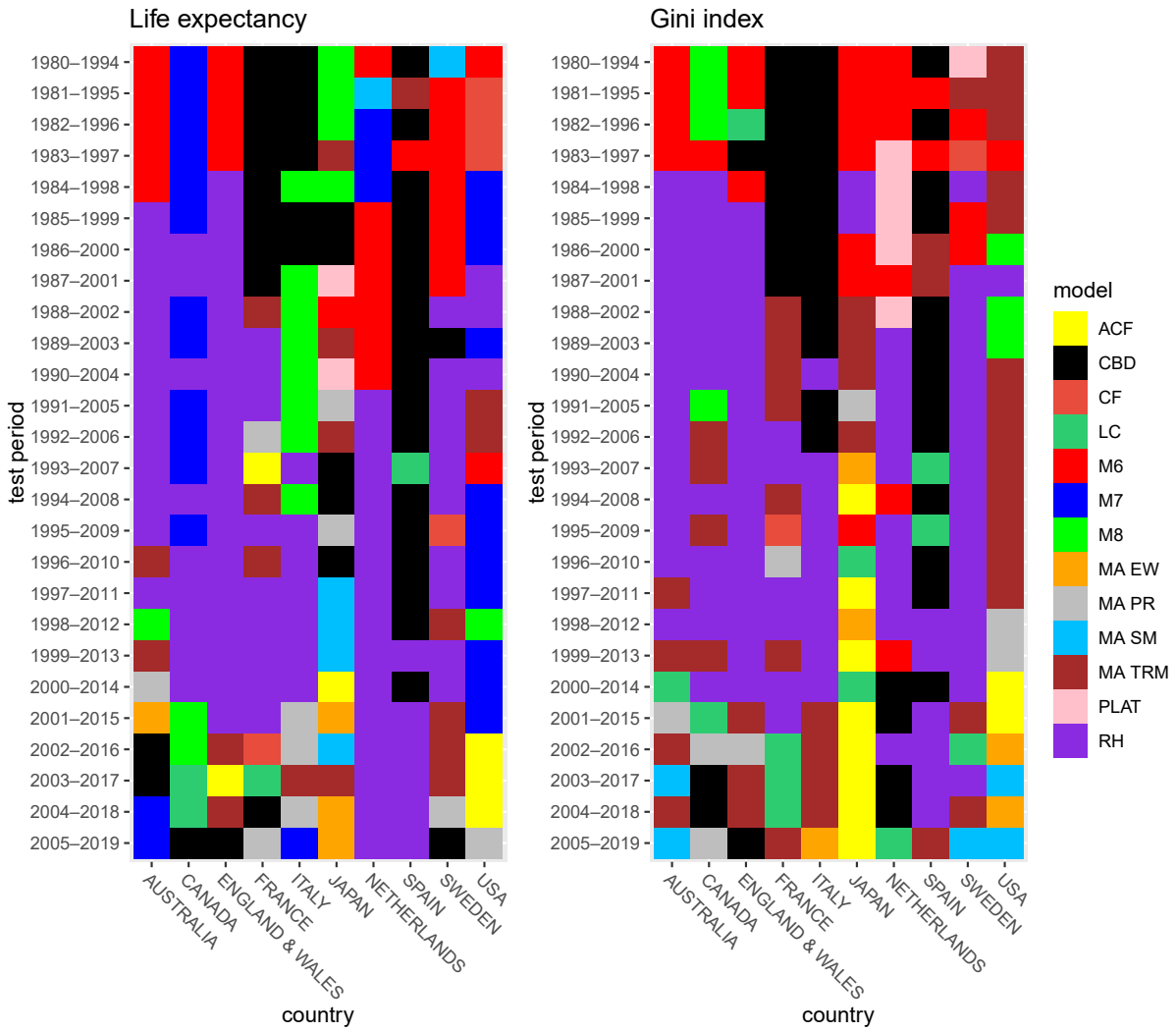


Figure 1.4: Model with the lowest MAFE by period and country. Results for individual models and model averaging approaches. Rolling test period case.

the last test period (see Table 1.2). This fact strengthens the motivation for choosing a model averaging approach that considers all or several models over a single model.
In Figure 1.4, and Tables 1.3 and 1.5, we observe how most individual models have a good performance in terms of interval forecast accuracy and number of cases with the lowest MAFE, for at least one country. But there are also countries where these models show poor performances, with low interval forecast accuracy and no cases in which they are the best. This is a consequence of the fact that each country has a specific mortality trend, which is fitted better by certain models than others. On the other hand, model averaging approaches have more robust results with good performances in all the countries.

1.6.2 Fixed test period

Observing Figure 1.5, we see how, similarly to what happened with the rolling test period case, the model averaging approaches based on equal and proportional weights, and on softmax function, have the best results in terms of MAFE. They all present a median MAFE lower than 0.2 and 0.004 respectively for life expectancy and the Gini index. Furthermore, notable performances in terms of median MAFE are obtained with the LC

Table 1.2: Interval forecast accuracy by period. Proportion of cases in which the observed life expectancy falls in the forecasting interval. Rolling test period case.

Training period	Test period	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR.W	MA S.M	MA TR
1950-1979	**1980-1994**	83%	84%	95%	84%	77%	95%	71%	82%	91%	100%	98%	100%	96%
1951-1980	**1981-1995**	82%	89%	95%	83%	81%	91%	71%	81%	86%	99%	98%	99%	96%
1952-1981	**1982-1996**	83%	87%	90%	79%	81%	88%	72%	81%	86%	99%	98%	99%	97%
1953-1982	**1983-1997**	81%	84%	91%	83%	79%	86%	81%	84%	84%	99%	98%	98%	96%
1954-1983	**1984-1998**	75%	81%	87%	75%	76%	82%	77%	77%	79%	98%	97%	97%	95%
1955-1984	**1985-1999**	86%	87%	92%	83%	84%	92%	87%	88%	86%	99%	98%	99%	97%
1956-1985	**1986-2000**	80%	82%	83%	77%	80%	78%	86%	75%	76%	98%	98%	98%	95%
1957-1986	**1987-2001**	81%	82%	85%	80%	80%	84%	88%	77%	76%	98%	97%	98%	96%
1958-1987	**1988-2002**	82%	83%	87%	77%	83%	82%	92%	82%	80%	98%	97%	98%	95%
1959-1988	**1989-2003**	79%	80%	76%	74%	81%	74%	90%	75%	75%	96%	95%	96%	92%
1960-1989	**1990-2004**	78%	84%	85%	76%	80%	80%	95%	77%	74%	97%	97%	97%	91%
1961-1990	**1991-2005**	86%	88%	79%	75%	89%	82%	89%	81%	81%	97%	97%	97%	96%
1962-1991	**1992-2006**	85%	89%	81%	75%	85%	82%	87%	79%	78%	96%	97%	97%	96%
1963-1992	**1993-2007**	86%	85%	79%	75%	83%	81%	84%	76%	82%	98%	97%	97%	90%
1964-1993	**1994-2008**	76%	79%	61%	59%	83%	63%	86%	66%	73%	91%	92%	92%	86%
1965-1994	**1995-2009**	84%	86%	70%	70%	82%	75%	87%	71%	81%	99%	99%	99%	95%
1966-1995	**1996-2010**	82%	85%	60%	60%	84%	65%	90%	70%	76%	97%	97%	97%	87%
1967-1996	**1997-2011**	81%	82%	61%	60%	81%	66%	90%	67%	76%	97%	98%	98%	96%
1968-1997	**1998-2012**	85%	81%	60%	62%	79%	72%	89%	70%	79%	98%	98%	98%	93%
1969-1998	**1999-2013**	82%	79%	53%	59%	74%	66%	88%	68%	76%	96%	95%	96%	89%
1970-1999	**2000-2014**	74%	75%	33%	41%	73%	54%	86%	61%	73%	94%	95%	95%	88%
1971-2000	**2001-2015**	88%	83%	45%	59%	77%	68%	72%	65%	80%	96%	94%	96%	91%
1972-2001	**2002-2016**	90%	86%	49%	70%	79%	68%	62%	64%	86%	98%	98%	98%	94%
1973-2002	**2003-2017**	87%	82%	35%	63%	74%	67%	70%	62%	85%	95%	94%	95%	78%
1974-2003	**2004-2018**	75%	82%	25%	52%	68%	56%	61%	59%	79%	92%	90%	92%	82%
1975-2004	**2005-2019**	90%	77%	51%	83%	77%	83%	41%	68%	87%	99%	99%	98%	95%
Average		82%	83%	70%	71%	80%	76%	81%	73%	80%	97%	97%	97%	93%

Table 1.3: Interval forecast accuracy by country. Proportion of cases in which the observed life expectancy falls in the forecasting interval. Rolling test period case.

	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR.W	MA S.M	MA TR
AUSTRALIA	90%	94%	82%	95%	96%	84%	91%	92%	91%	99%	99%	99%	97%
CANADA	61%	62%	49%	78%	59%	54%	58%	62%	52%	99%	96%	99%	78%
ENGLAND & WALES	74%	75%	78%	69%	71%	70%	89%	63%	75%	99%	99%	99%	99%
FRANCE	97%	96%	34%	63%	89%	80%	88%	87%	97%	100%	99%	100%	97%
ITALY	84%	88%	70%	59%	88%	76%	83%	88%	85%	97%	97%	98%	95%
JAPAN	99%	87%	87%	62%	75%	96%	80%	50%	99%	100%	100%	100%	96%
NETHERLANDS	56%	61%	77%	71%	57%	76%	80%	67%	58%	86%	87%	87%	86%
SPAIN	98%	98%	85%	70%	95%	94%	95%	88%	99%	100%	100%	100%	95%
SWEDEN	73%	80%	62%	42%	78%	65%	71%	69%	66%	91%	88%	90%	84%
USA	91%	92%	71%	96%	89%	67%	71%	67%	80%	100%	100%	100%	98%
Average	82%	83%	70%	71%	80%	76%	81%	73%	80%	97%	97%	97%	93%

Table 1.4: Interval forecast accuracy by period. Proportion of cases in which the observed Gini index falls in the forecasting interval. Rolling test period case.

Training period	Test period	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR	MA S.M	MA TR
1950-1979	**1980-1994**	76%	73%	92%	71%	77%	69%	53%	55%	69%	94%	94%	94%	91%
1951-1980	**1981-1995**	75%	74%	92%	71%	76%	57%	53%	54%	65%	92%	93%	92%	91%
1952-1981	**1982-1996**	77%	75%	89%	70%	77%	48%	51%	53%	66%	91%	91%	91%	90%
1953-1982	**1983-1997**	73%	72%	85%	69%	78%	50%	63%	51%	64%	88%	90%	89%	85%
1954-1983	**1984-1998**	70%	69%	82%	64%	74%	45%	63%	46%	63%	89%	90%	88%	87%
1955-1984	**1985-1999**	78%	73%	88%	62%	81%	53%	74%	51%	68%	93%	93%	93%	92%
1956-1985	**1986-2000**	68%	75%	80%	54%	75%	38%	65%	47%	63%	87%	87%	87%	87%
1957-1986	**1987-2001**	68%	67%	82%	62%	78%	49%	77%	50%	62%	88%	89%	89%	89%
1958-1987	**1988-2002**	68%	65%	82%	57%	76%	42%	87%	54%	62%	87%	86%	86%	86%
1959-1988	**1989-2003**	69%	68%	84%	58%	76%	37%	88%	52%	62%	90%	91%	90%	85%
1960-1989	**1990-2004**	64%	67%	83%	60%	77%	46%	89%	58%	62%	87%	88%	87%	89%
1961-1990	**1991-2005**	80%	75%	81%	58%	83%	37%	89%	57%	68%	90%	91%	90%	92%
1962-1991	**1992-2006**	79%	74%	87%	60%	80%	46%	84%	58%	68%	92%	92%	91%	91%
1963-1992	**1993-2007**	85%	73%	88%	60%	81%	44%	88%	57%	77%	93%	93%	93%	92%
1964-1993	**1994-2008**	70%	76%	73%	53%	76%	20%	85%	56%	71%	88%	92%	89%	94%
1965-1994	**1995-2009**	81%	71%	77%	58%	72%	42%	88%	57%	77%	96%	96%	96%	97%
1966-1995	**1996-2010**	75%	78%	73%	55%	75%	29%	89%	57%	78%	93%	96%	92%	91%
1967-1996	**1997-2011**	77%	78%	75%	55%	75%	33%	87%	58%	79%	92%	95%	91%	97%
1968-1997	**1998-2012**	80%	73%	74%	51%	69%	41%	89%	59%	82%	94%	95%	94%	97%
1969-1998	**1999-2013**	81%	72%	67%	51%	71%	33%	84%	56%	83%	92%	94%	92%	92%
1970-1999	**2000-2014**	77%	75%	49%	42%	65%	12%	83%	56%	80%	90%	94%	91%	84%
1971-2000	**2001-2015**	88%	81%	57%	55%	71%	27%	72%	61%	82%	96%	96%	96%	94%
1972-2001	**2002-2016**	92%	78%	64%	59%	75%	35%	62%	62%	87%	99%	100%	99%	96%
1973-2002	**2003-2017**	90%	75%	48%	56%	70%	28%	71%	64%	89%	96%	95%	96%	89%
1974-2003	**2004-2018**	82%	83%	40%	50%	69%	27%	64%	60%	89%	94%	94%	94%	93%
1975-2004	**2005-2019**	86%	68%	60%	72%	78%	46%	42%	70%	87%	100%	98%	100%	94%
Average		77%	73%	75%	59%	75%	40%	75%	56%	73%	92%	93%	92%	91%

Table 1.5: Interval forecast accuracy by country. Proportion of cases in which the observed Gini index falls in the forecasting interval. Rolling test period case.

	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA P.E	MA S.M	MA TR
AUSTRALIA	84%	88%	94%	85%	87%	32%	94%	73%	79%	94%	95%	94%	98%
CANADA	55%	61%	60%	90%	57%	14%	58%	34%	54%	84%	88%	84%	90%
ENGLAND & WALES	58%	62%	93%	63%	52%	34%	76%	48%	68%	88%	92%	88%	92%
FRANCE	97%	82%	39%	51%	89%	36%	89%	83%	97%	100%	100%	100%	99%
ITALY	69%	83%	67%	41%	73%	31%	75%	61%	62%	87%	88%	87%	86%
JAPAN	90%	73%	92%	39%	84%	59%	82%	52%	96%	100%	100%	100%	92%
NETHERLANDS	63%	45%	82%	58%	62%	51%	55%	34%	50%	81%	84%	81%	81%
SPAIN	99%	96%	88%	55%	94%	51%	95%	84%	99%	100%	100%	100%	100%
SWEDEN	66%	69%	53%	16%	71%	46%	61%	37%	53%	82%	81%	82%	73%
USA	87%	80%	88%	86%	82%	43%	74%	50%	68%	100%	100%	100%	97%
Average	77%	73%	75%	59%	75%	40%	75%	56%	73%	92%	93%	92%	91%

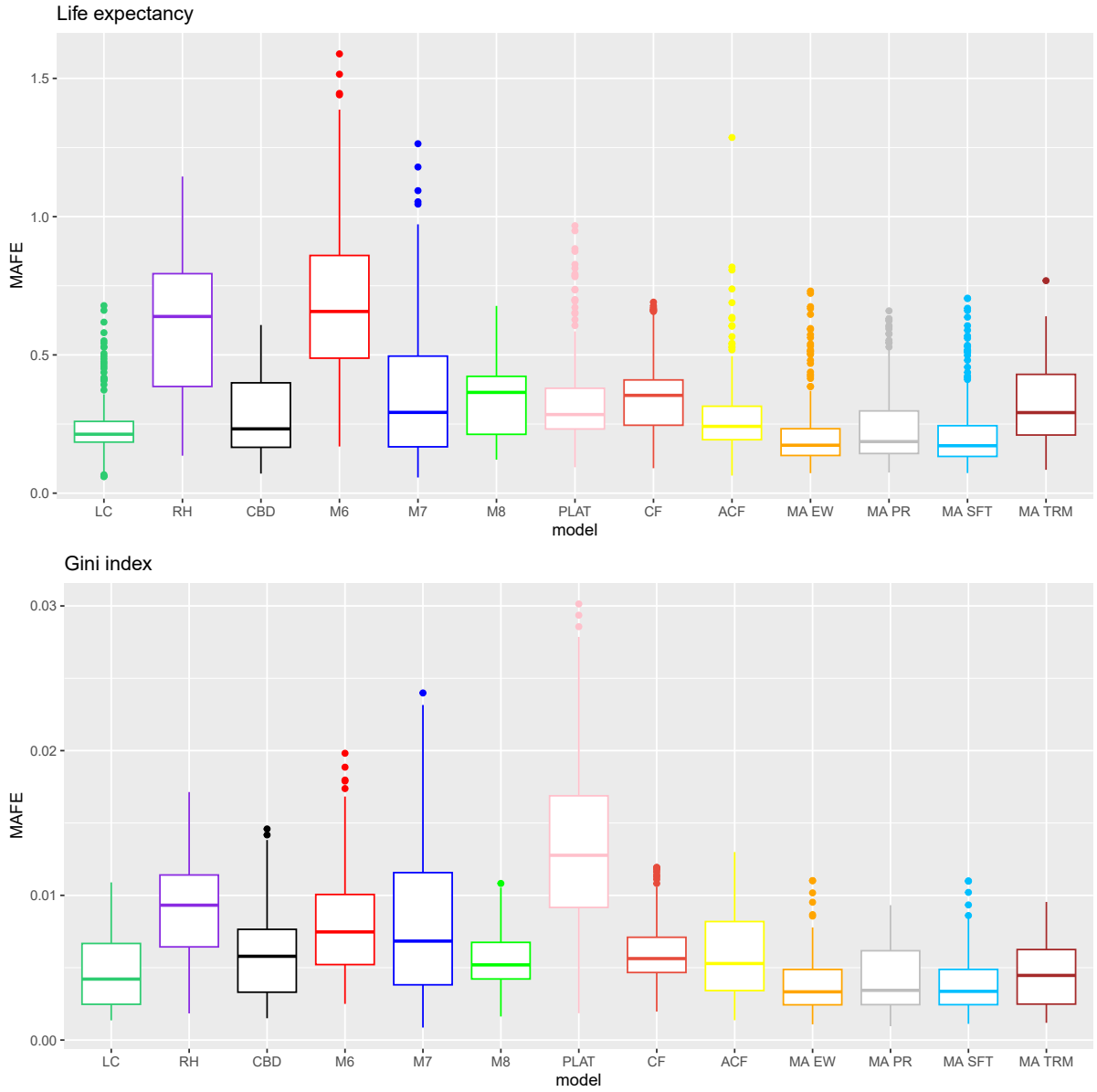


Figure 1.5: Summary of the MAFEs by model. Results for individual models and according weighted average. Fixed test period case.

model for both the Gini index and life expectancy, while here the RH model is among the worst for both life expectancy and the Gini index. The model averaging approach based on trimming has a higher MAFE and it is outperformed by other models such as CBD and ACF for life expectancy, and LC for the Gini index. Indeed, the trimming based averaging approach heavily relies on the RH model which has poor performances in the test period considered.

Figure 1.6 shows the model with the lowest MAFE by period and country. The RH model still results in having the highest number of best performances for life expectancy (16%),[6] while the LC model results in having the highest number for the Gini index

[6]These percentages have been calculated as the ratio between the number of cases in which each model is the best over the total number of cases considered (200).

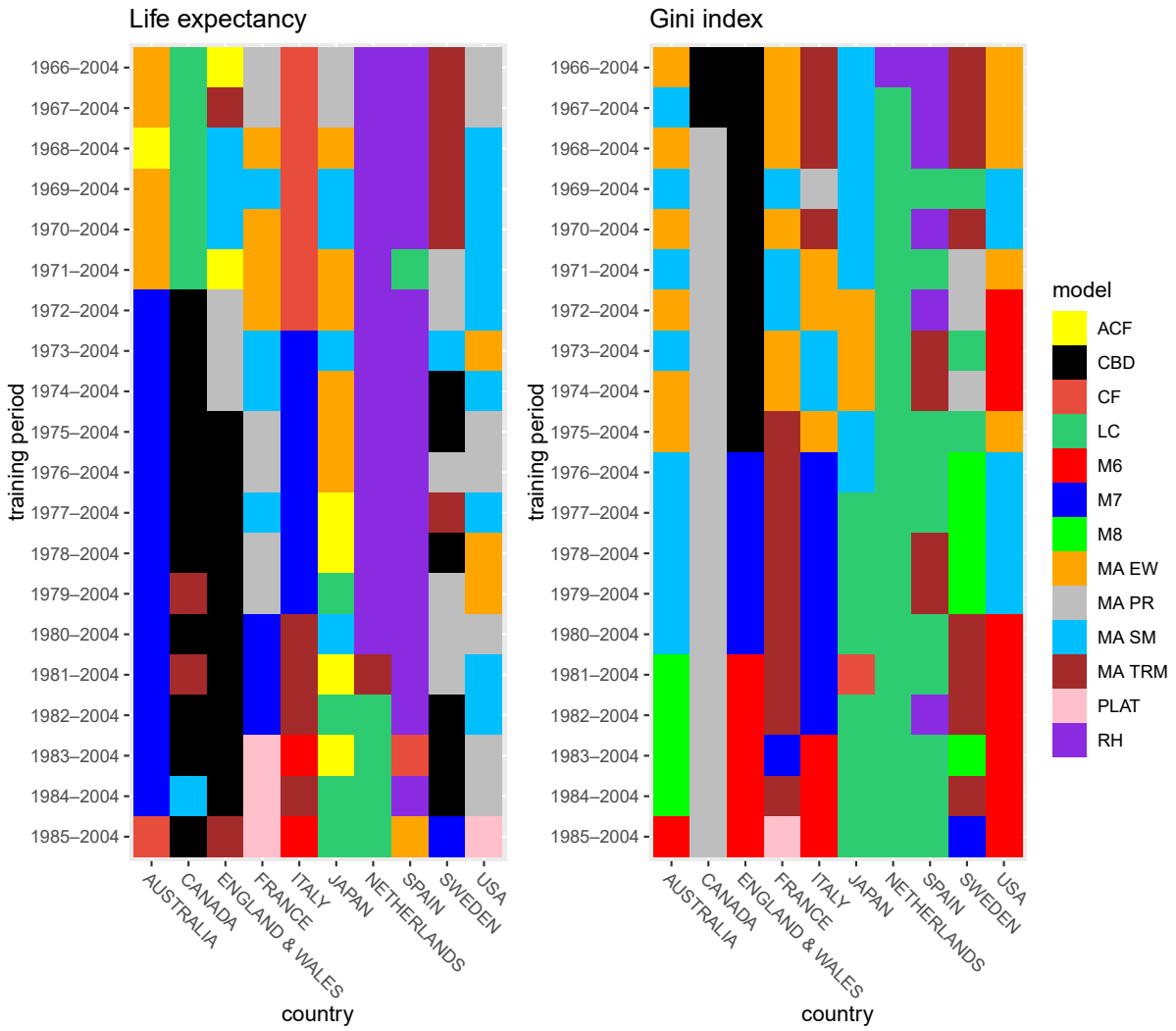


Figure 1.6: Model with the lowest MAFE by period and country. Results for individual models and model averaging approaches. Fixed test period case.

(20%). All the model averaging approaches here have good performances, indeed they globally account for 40% for life expectancy and 49% for the Gini index. Other remarkable results are obtained with the CBD and M7 models for life expectancy, and the M6 model for what concerns the Gini index. As in the rolling period case, the apparent discrepancy between some results in Figure 1.5 and those in Figure 1.6 is explained by the superior robustness of the averaging approaches over the individual models.
Tables 1.6-1.9 show the proportion of cases the observed metrics fall in the forecasting intervals. As was the case for the rolling test period case, the model averaging approaches based on equal and proportional weights, and on softmax function have the highest interval forecast accuracy with 98% for life expectancy and 99% for the Gini index. They are followed by the model averaging approach based on trimming (94% and 95%). Among the other models, we find that the LC and ACF models show good results for life expectancy as well as for the Gini index. The RH model here is the worst model for both life expectancy and the Gini index.

To conclude, in Tables 1.6-1.9 and Figures 1.5 and 1.6, we notice that, compared with the

rolling test period case, there is less variability in the models' performances by training period as the test period remains fixed. A similar observation can be made regarding the results for life expectancy and the Gini index. Instead, for what concerns the results by country, we observe how the performances' variability remains, with many individual models performing well in some countries and badly in others. In contrast, the model averaging approaches, except for the one based on trimming that is slightly weaker, show good results in all the countries considered.

Table 1.6: Interval forecast accuracy by period. Proportion of cases in which the observed life expectancy falls in the forecasting interval. Fixed test period case.

Training period	Test period	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR.W	MA S.M	MA TR
1966-2004	**2005-2019**	91%	84%	48%	62%	83%	78%	62%	69%	84%	98%	99%	97%	93%
1967-2004	**2005-2019**	91%	84%	49%	63%	82%	77%	59%	70%	85%	97%	99%	98%	93%
1968-2004	**2005-2019**	91%	82%	52%	67%	80%	79%	55%	71%	88%	96%	97%	97%	92%
1969-2004	**2005-2019**	90%	81%	52%	72%	79%	80%	54%	71%	88%	97%	99%	99%	93%
1970-2004	**2005-2019**	91%	81%	50%	70%	79%	79%	50%	70%	88%	98%	99%	98%	94%
1971-2004	**2005-2019**	91%	80%	49%	75%	79%	79%	47%	69%	88%	99%	99%	99%	94%
1972-2004	**2005-2019**	91%	80%	47%	75%	78%	78%	45%	66%	87%	99%	99%	99%	93%
1973-2004	**2005-2019**	91%	78%	50%	78%	77%	81%	43%	68%	88%	98%	99%	98%	89%
1974-2004	**2005-2019**	91%	79%	51%	81%	78%	82%	44%	68%	88%	99%	98%	98%	90%
1975-2004	**2005-2019**	90%	77%	50%	83%	78%	84%	40%	69%	87%	97%	99%	99%	94%
1976-2004	**2005-2019**	90%	78%	52%	86%	78%	85%	39%	70%	88%	99%	98%	98%	95%
1977-2004	**2005-2019**	91%	79%	49%	85%	77%	84%	40%	72%	89%	96%	98%	97%	94%
1978-2004	**2005-2019**	90%	79%	51%	87%	78%	85%	37%	75%	90%	97%	98%	97%	96%
1979-2004	**2005-2019**	90%	81%	49%	89%	78%	84%	37%	76%	90%	97%	98%	98%	94%
1980-2004	**2005-2019**	89%	79%	52%	89%	77%	85%	35%	73%	88%	98%	98%	97%	99%
1981-2004	**2005-2019**	89%	79%	53%	91%	77%	85%	35%	74%	88%	98%	98%	98%	100%
1982-2004	**2005-2019**	89%	80%	56%	90%	78%	84%	36%	75%	88%	98%	98%	98%	98%
1983-2004	**2005-2019**	86%	80%	57%	90%	77%	82%	33%	76%	86%	99%	99%	99%	95%
1984-2004	**2005-2019**	86%	82%	57%	93%	78%	86%	34%	76%	87%	99%	99%	99%	97%
1985-2004	**2005-2019**	82%	78%	62%	93%	80%	85%	34%	75%	85%	99%	99%	99%	94%
Average		90%	80%	52%	81%	79%	82%	43%	72%	87%	98%	98%	98%	94%

Table 1.7: Interval forecast accuracy by country. Proportion of cases in which the observed life expectancy falls in the forecasting interval. Fixed test period case.

	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR.W	MA S.M	MA TR
Australia	96%	100%	71%	100%	100%	98%	37%	95%	99%	100%	100%	100%	97%
Canada	85%	87%	28%	94%	84%	52%	30%	32%	65%	99%	98%	99%	84%
England & Wales	99%	100%	66%	100%	98%	79%	32%	64%	96%	100%	100%	100%	100%
France	90%	54%	18%	89%	64%	97%	33%	98%	92%	100%	100%	100%	99%
Italy	91%	73%	87%	96%	69%	93%	24%	97%	95%	100%	100%	100%	83%
Japan	99%	48%	58%	73%	50%	100%	53%	60%	100%	100%	100%	100%	100%
Netherlands	55%	62%	47%	45%	58%	50%	80%	48%	51%	80%	87%	82%	97%
Spain	100%	90%	39%	65%	77%	93%	99%	84%	100%	100%	100%	100%	92%
Sweden	100%	100%	52%	81%	100%	89%	21%	93%	99%	100%	100%	100%	100%
USA	82%	88%	48%	66%	87%	71%	21%	46%	79%	100%	100%	100%	92%
Average	90%	80%	52%	81%	79%	82%	43%	72%	87%	98%	98%	98%	94%

Table 1.8: Interval forecast accuracy by period. Proportion of cases in which the observed Gini index falls in the forecasting interval. Fixed test period case.

Training period	Test period	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR	MA S.M	MA TR
1966-2004	**2005-2019**	93%	77%	48%	49%	79%	32%	62%	65%	76%	97%	98%	97%	97%
1967-2004	**2005-2019**	93%	77%	51%	50%	79%	33%	60%	65%	78%	97%	98%	97%	97%
1968-2004	**2005-2019**	91%	72%	54%	53%	78%	39%	54%	66%	84%	98%	98%	98%	96%
1969-2004	**2005-2019**	91%	70%	58%	57%	79%	40%	56%	68%	84%	98%	98%	98%	95%
1970-2004	**2005-2019**	91%	73%	57%	52%	78%	43%	50%	66%	85%	99%	98%	99%	96%
1971-2004	**2005-2019**	92%	71%	54%	56%	77%	43%	47%	66%	87%	99%	98%	99%	96%
1972-2004	**2005-2019**	90%	71%	53%	57%	78%	42%	46%	68%	87%	100%	98%	100%	95%
1973-2004	**2005-2019**	89%	67%	55%	60%	77%	44%	43%	70%	87%	99%	98%	99%	94%
1974-2004	**2005-2019**	86%	67%	58%	66%	79%	45%	43%	70%	87%	99%	98%	99%	95%
1975-2004	**2005-2019**	86%	68%	59%	73%	78%	47%	42%	71%	87%	100%	98%	100%	93%
1976-2004	**2005-2019**	83%	68%	64%	74%	79%	50%	40%	68%	87%	100%	98%	100%	91%
1977-2004	**2005-2019**	84%	68%	60%	74%	78%	48%	40%	72%	88%	99%	98%	99%	91%
1978-2004	**2005-2019**	83%	68%	63%	75%	78%	50%	36%	70%	86%	100%	98%	100%	92%
1979-2004	**2005-2019**	84%	69%	63%	77%	78%	50%	36%	73%	86%	99%	99%	99%	95%
1980-2004	**2005-2019**	82%	68%	64%	76%	78%	52%	35%	70%	84%	100%	99%	100%	98%
1981-2004	**2005-2019**	82%	68%	65%	80%	79%	55%	34%	71%	84%	100%	99%	100%	98%
1982-2004	**2005-2019**	80%	68%	67%	78%	80%	55%	35%	73%	80%	100%	98%	100%	97%
1983-2004	**2005-2019**	76%	65%	69%	79%	79%	56%	33%	74%	78%	100%	99%	100%	97%
1984-2004	**2005-2019**	75%	66%	74%	82%	81%	57%	36%	74%	80%	100%	98%	100%	96%
1985-2004	**2005-2019**	70%	61%	78%	80%	79%	60%	34%	69%	74%	100%	98%	100%	94%
Average		85%	69%	61%	67%	79%	47%	43%	69%	83%	99%	98%	99%	95%

Table 1.9: Interval forecast accuracy by country. Proportion of cases in which the observed Gini index falls in the forecasting interval. Fixed test period case.

	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA P.E	MA S.M	MA TR
AUSTRALIA	85%	93%	93%	88%	95%	57%	36%	90%	87%	100%	99%	100%	93%
CANADA	69%	95%	51%	92%	89%	22%	53%	29%	88%	100%	100%	100%	100%
ENGLAND & WALES	94%	98%	92%	99%	77%	34%	38%	86%	80%	100%	100%	100%	100%
FRANCE	79%	5%	41%	79%	81%	67%	16%	89%	84%	100%	100%	100%	95%
ITALY	91%	60%	88%	88%	88%	56%	28%	94%	90%	100%	100%	100%	97%
JAPAN	97%	47%	50%	49%	70%	94%	50%	74%	99%	100%	100%	100%	100%
NETHERLANDS	100%	87%	33%	34%	49%	35%	67%	42%	81%	92%	98%	92%	100%
SPAIN	100%	71%	42%	47%	87%	33%	99%	85%	100%	100%	100%	100%	100%
SWEDEN	98%	94%	54%	72%	100%	50%	27%	66%	90%	100%	100%	100%	100%
USA	38%	41%	64%	27%	50%	26%	17%	40%	36%	100%	85%	100%	68%
Average	85%	69%	61%	67%	79%	47%	43%	69%	83%	99%	98%	99%	95%

1.7 Conclusion

In this paper, we compared the forecasting performance of existing two-population models with four different model averaging approaches. We considered ten countries and 46 combinations of training and test periods, using truncated life expectancy and the Gini index as evaluation metrics. Our results show that model averaging approaches generally outperformed the individual models, achieving superior results, with the only exception of the RH model in the rolling test period case, both in terms of MAFE (difference between the forecasted and observed truncated life expectancy and Gini index) and interval forecast accuracy (proportion of cases in which observed values fall within the prediction interval). Among the model averaging approaches, the best results were obtained with those using equal, proportional, or softmax-based weights, while the trimming-based approach performed slightly worse, although it was still better than many individual models. The higher MAFE of the trimming approach is likely due to changes in mortality trends: models that perform well in the validation period do not necessarily perform equally well in the test period because of accelerations or decelerations in mortality improvements over time. Finally, a further advantage of model averaging approaches is that, whereas the performance of individual models is affected by the choice of metric (life expectancy or Gini index), country, and period, model averaging approaches are shown to be more robust in this regard. For future research, a natural extension would be to consider multi-population models that simultaneously include three or more populations. Furthermore, alternative evaluation metrics beyond MAFE and interval forecast accuracy could be explored to compare the effectiveness of model averaging approaches with traditional models.

Chapter 2

Mortality forecasting via multi-task neural networks

2.1 Introduction

Artificial neural networks, abbreviated as neural networks (NNs), are a subfield of machine learning, commonly referred to as deep learning, that have been applied to demography in recent years for analyzing and predicting mortality rates and other mortality-related metrics. Generally speaking, an NN can be seen as a universal function approximator, i.e., a mapping that, once properly structured and trained, can approximate any function that links a series of inputs to outputs, see Hornik et al. (1989). Focusing on mortality forecasting, there are two advantages of using NNs instead of traditional stochastic models such as the Lee-Carter Model and its extensions, see Lee and Carter (1992). Firstly, they simplify the model definition and free us from specifying how variables, such as age and calendar year, interact. Secondly, they allow us to consider the mortality experience of several populations simultaneously. Among the most important contributions to NNs applied to mortality forecasting, the following studies are among the ones that stand out: Richman and Wüthrich (2021), and Perla and Scognamiglio (2023) exploit feedforward NNs; Nigri et al. (2019), Chen and Khaliq (2022), Lindholm and Palmborg (2022), and Euthum et al. (2024) use long short-term memory NNs; Perla et al. (2021), Wang et al. (2021), and Schnürch and Korn (2022) utilize convolutional NNs; and Hainaut (2018) as well as Scognamiglio (2022) apply hybrid models.

In this paper we focus on simultaneously forecasting the mortality rates of a given set of countries. In order to do that, we implement a methodology called multi-task NNs, consisting of several NNs that share a certain number of parameters. In the past years, multi-task deep learning has been applied with promising results in several fields, such as computer vision, see Girshick (2015), natural language processing, see Collobert and Weston (2008), speech recognition, see Deng et al. (2013), and insurance, see Lindholm et al. (2023). Finally, we recommend Zhang and Yang (2021) for a theoretical overview of multi-task NNs.

Specifically, we propose a hierarchical network structure for multi-population mortality forecasting. The lower hidden layers of these multi-task NNs, i.e. those closer to the input layer, are shared across all countries, capturing the general properties of mortality trends, while the higher hidden layers, i.e. those closer to the output layer, are country-specific or shared only within clusters of countries with more similar past mortality trends. The clusters are obtained by applying the k-means clustering machine learning technique to

past data for some key mortality metrics, i.e. life expectancy and lifetime standard deviation. Finally, each country has its own layer to learn its distinct property.
In this paper, we quantitatively compare multi-task NNs with pre-existing single-task NNs and stochastic models considering mortality data of seventeen different countries. The comparison is based of mortality rates, life expectancy and lifetime standard deviation forecasting errors. With multi-task NNs, we expect to improve the performance of NNs at country-specific level dedicating more parameters to single countries.
Our main conclusions are that multi-task NNs performance compared to single-task NNs and stochastic models depends on the metric, age range, and training period considered. Overall, single-task NNs gives the best results in terms of mortality rates forecasting error, while multi-task NNs and stochastic models have the lowest forecasting error respectively for life expectancy and lifetime standard deviation. Furthermore, implementing a weighting scheme in their training improves the multi-task NNs performance, especially for life expectancy and lifetime standard deviation when considering wider age ranges.
The remainder of this paper is organized as follows. Section 2.2 contains a general theoretical framework for feedforward NNs, followed by a practical application to mortality rates forecasting. In Section 2.3, we introduce feedforward multi-task NNs and present the NNs proposed by us. In Section 2.4, the data used in the empirical analysis and settings for the training of the NNs are reported. In Section 2.5, the numerical results are presented and discussed. In Section 2.6, we draw the conclusion and propose some future outlooks.

2.2 Feedforward neural networks

Feedforward neural networks (FNNs) are the most basic type of NN. Information flows in one direction, from input neurons through hidden layers to output neurons, see Schmidhuber (2015). Cycles and loops are not present in this type of NN. They are generally used for classification, regression, and pattern recognition, and, in particular, they can be applied to mortality forecasting. In this context, FNNs are especially useful for mortality forecasting when the focus is on modelling the relationship between input features (age, calendar year, cohort year, etc.) and mortality rates.

2.2.1 Notation and terminology

Given a set of L input variables $\mathbf{X} = (X_1, \ldots, X_L)$ that can be numerical or categorical, or a combination of them, and the corresponding output Y, we have to focus on the hyperparameters of the NNs, i.e. those settings that have to be set before the parameters are learnt in the training process, see Goldberg (2017) and Prince (2023). These hyperparameters are:

- N: number of hidden layers in the NN.
- L_1,...,L_N: numbers of neurons for each layer.
- $f^{(1)}$,..., $f^{(N+1)}$: activation functions of the NN. Notice: $f^{(1)}$ will be the activation function of the first hidden layer, while $f^{(N+1)}$ will be the activation function of the output layer. Some popular activation functions, that are also used in this paper, are Sigmoid (also called Logistic), Hyperbolic Tangent (tanh), and Rectified Linear Unit (ReLU), see Dubey et al. (2022).

Once we have specified these hyperparameters, it is possible to estimate the parameters $\mathbf{B}^{(1)} \in \mathbb{R}^{L_1 \times L}$, $\mathbf{B}^{(2)} \in \mathbb{R}^{L_2 \times L_1}$, $\ldots$, $\mathbf{B}^{(N)} \in \mathbb{R}^{L_N \times L_{N-1}}$, $\mathbf{B}^{(N+1)} \in \mathbb{R}^{1 \times L_N}$, and $\mathbf{c}_1 \in \mathbb{R}^{L_1}$, $\mathbf{c}_2 \in \mathbb{R}^{L_2}$, $\ldots$, $\mathbf{c}_N \in \mathbb{R}^{L_N}$, $c_{N+1} \in \mathbb{R}$, that represent respectively weight matrices and intercept vectors. These are the parameters that are learned during the training of the network.
The layers will be so computed, using matrix notation:

$$\mathbf{Z}^{(1)} = f^{(1)}(\mathbf{c}_1 + \mathbf{B}^{(1)}\mathbf{X}) \in \mathbb{R}^{L_1}, \tag{2.1}$$

where $\mathbf{X} \in \mathbb{R}^L$ is the input vector,

$$\mathbf{Z}^{(j)} = f^{(j)}(\mathbf{c}_j + \mathbf{B}^{(j)}\mathbf{Z}^{(j-1)}) \in \mathbb{R}^{L_j},\ j = 2, \ldots, N. \tag{2.2}$$

Finally, for the output layer:

$$\hat{Y} = Z^{(N+1)} = f^{(N+1)}(c_{N+1} + \mathbf{B}^{(N+1)}\mathbf{Z}^{(N)}) \in \mathbb{R}. \tag{2.3}$$

We now discuss the training of the NN during which all the weight matrices and intercept vectors are estimated through a process called backpropagation. In order to do that, the additional hyperparameters reported below have to be specified, see Prince (2023).

- The **loss function** is the criterion through which, starting from the observed value of the outputs and the predicted output of the network, we calculate the quantity that has to be minimized when we train the NN. In the remainder of this paper, the mean squared error (MSE) is used as loss function:

$$MSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2, \tag{2.4}$$

  where n is the number of observations, Y_i are the observed values of the output, $\hat{Y}_i$ are the values predicted by the NN as in equation (3.7).

- The **optimizer** is the algorithm used during the training phase to adjust the parameters of the neural network in order to minimize the loss. In the remainder of this paper, we will utilize the Adam optimizer (Adaptive Moment Estimation), a gradient-based optimization algorithm that leverages first-order (gradient) and second-order (squared gradient) moment estimates to adapt the learning rate for each parameter, see Kingma and Ba (2014).

- The **number of epochs** is the amount of times the optimizer runs on the training set.

- The **validation set** is a subset of the available data used to provide an unbiased evaluation of a model fit identifying eventual overfitting while the training set is used to tune the NN parameters.

- The **batch size** defines the number of training samples processed simultaneously before the model's weights are updated. It determines how many samples are passed through the network in each forward and backward pass during training.

- The **learning rate** controls the size of the steps taken during the optimization process when adjusting the weights of the model.

2.2.2 Feedforward single-task neural networks applied to mortality forecasting

In this subsection, we are going to provide a framework for forecasting of mortality rates with feedforward single-task NNs based on the paper of Richman and Wüthrich (2021). The input variables considered in the NNs are calendar year t, age x, gender g, and country p, $\tilde{\mathbf{X}} = (t, x, g, p)$, and they will be treated as categorical with the single exception of calendar year, which will be treated as numerical, while the output, Y, is the central mortality rate $m_{x,t}^{(g,p)}$ at age x, year t, gender g and population p. In order to treat the categorical variables in the input layer, embedding layers are used, see Mikolov et al. (2013). An embedding layer, from a mathematical point of view, is a function that maps discrete data into continuous vector representations. So, given a categorical variable with b distinct categories or levels (e.g., the categories "male" and "female" for the variable "gender", the different countries for the variable "country", etc.), and a dimension d, which represents the size of the continuous embedding space (e.g., each categorical level will be represented by a vector in $\mathbb{R}^d$), the embedding layer performs the mapping

$$f : \{0, 1, \dots, b-1\} \to \mathbb{R}^d. \tag{2.5}$$

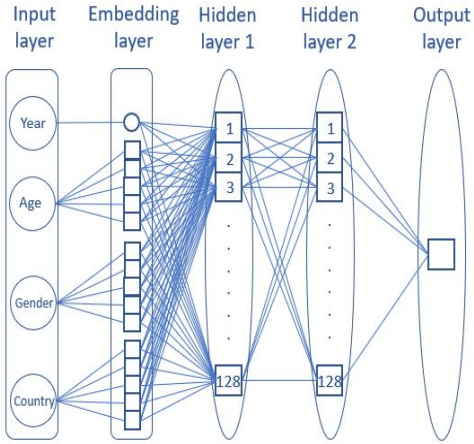
In a NN, the embedding layer can be identified with a parametrized matrix belonging to $\mathbb{R}^{b\times d}$. The parameters of the embedding layers, similarly to the parameters of the hidden layers, are learned as the network is trained. Notice that if $d = 1$, the embedding layer becomes equivalent to the classical treatment of categorical variables in regression models: each level of the variable is coded with a specific value. Following Richman and Wüthrich (2021), d is set equal to 5 for all three categorical variables, so: $x \to \mathbf{x} \in \mathbb{R}^5$, $g \to \mathbf{g} \in \mathbb{R}^5$, and $p \to \mathbf{p} \in \mathbb{R}^5$. Once embedding vectors ($\mathbf{x}$, $\mathbf{g}$ and $\mathbf{p}$) have been created, we have the vector $\mathbf{X} = (t, \mathbf{x}, \mathbf{g}, \mathbf{p}) \in \mathbb{R}^{16}$ that represents the actual input that will be passed to the first hidden layer of the NN. The number of hidden layers here considered differs by the NN considered, $N = 2$ or 5; the number of neurons in each hidden layer is equal to 128 neurons, $L_1 = \cdots = L_N = 128$; the output layer that represents the mortality rate for the gender g in the country p at age x in year t has one neuron, $L_{N+1} = 1$, with sigmoid activation function,

$$m_{x,t}^{(g,p)} = Z^{(N+1)} = \frac{1}{1 + e^{-(c_{N+1} + \mathbf{B}^{(N+1)}\mathbf{Z}^{(N)})}}. \tag{2.6}$$

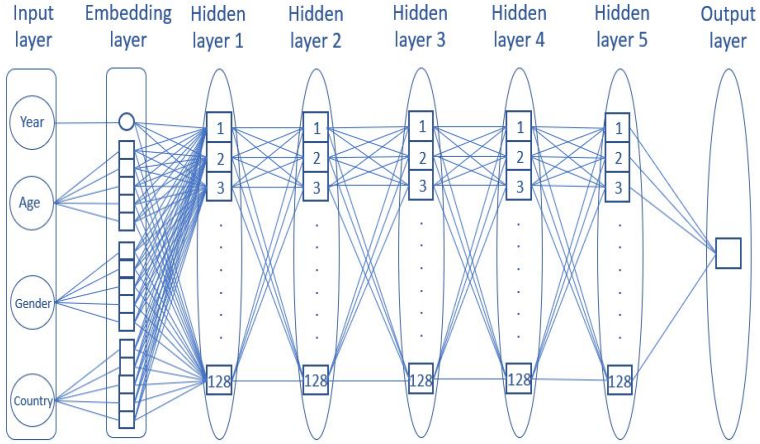
The NNs also differ among themselves by the type of activation function in the hidden layers, $f^{(1)} = \cdots = f^{(N)} = \tanh$ or $f^{(1)} = \cdots = f^{(N)} =$ReLU, and by the presence or not of a direct connection, called a skip connection, between the embedding layer and the last hidden layer. These NNs are referred as DEEPi, $i = 1, \dots, 6$, and the details about their architecture are reported in Table 2.1 and in Figure 2.1.

Table 2.1: Summary of the NNs DEEPi, $i = 1, \dots, 6$, architectures.

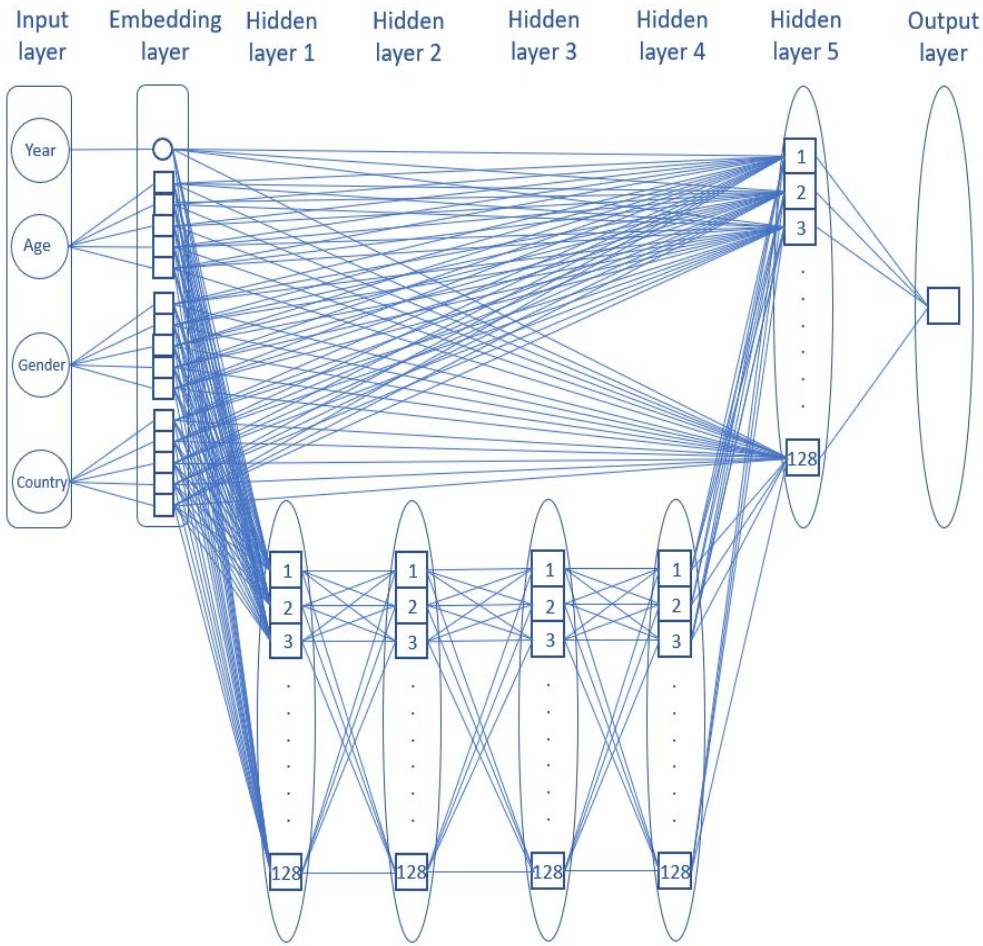
model	# hidden layers	activation function	skip connection
DEEP1	2	ReLU	No
DEEP2	2	tanh	No
DEEP3	5	ReLU	No
DEEP4	5	tanh	No
DEEP5	5	ReLU	Yes
DEEP6	5	tanh	Yes



(a) DEEP1 and DEEP2



(b) DEEP3 and DEEP4



(c) DEEP5 and DEEP6

Figure 2.1: Architecture of the NNs DEEPi, $i = 1, \ldots, 6$ as described in Table 2.1.

2.3 Multi-task neural networks

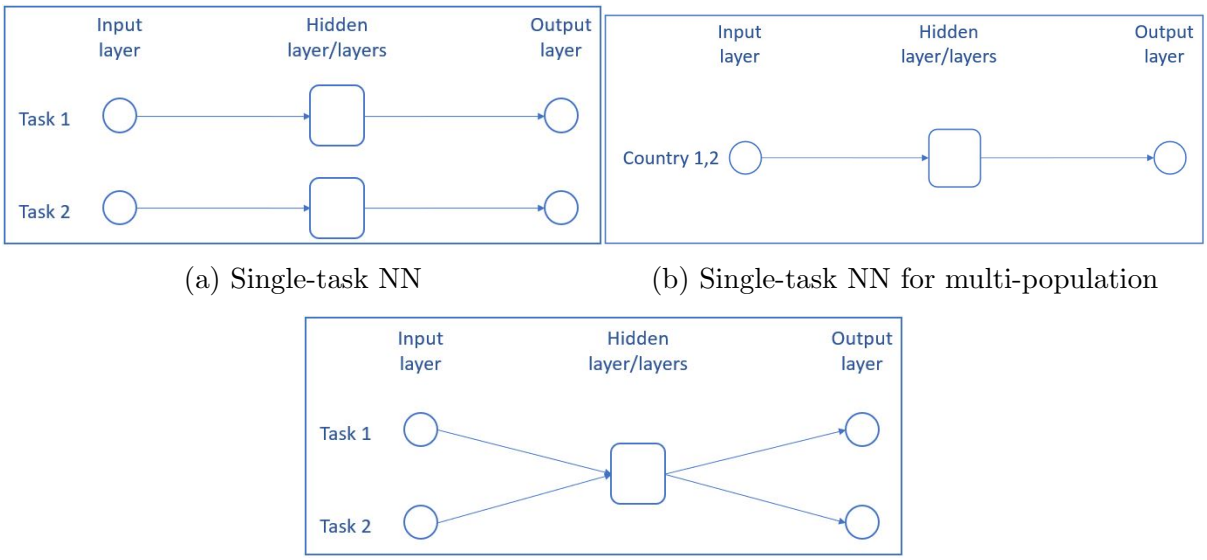
Generally speaking, multi-task deep learning consists of different NNs (one for each task) that share at least one layer. The shared part of the NNs can be the input layer, one or more hidden layers, or a combination of them. It is relevant to notice that the output layer cannot be shared, as we must have one output neuron for each task. Given that we have P different datasets, each corresponding to a distinct country, this paper employs multi-task NNs with a multi-input, multi-output structure, see Menet et al. (2023). Specifically, these NNs share hidden layers across tasks while maintaining P separate input and output layers.

Let us now consider $P > 1$ countries and the following P tasks: T_p = "forecasting the mortality rates for p^{th} country", $p = 1, \ldots, P$. If we want to forecast the mortality rates of the P countries using a feed-forward NN, then we have three different options. The first consists of using P different NNs with their own input layer, hidden layers, and output layer, with the p^{th} of them to predict the mortality rates of the p^{th} country. This solution can be called single-task NNs approach and is graphically represented in Figure 2.2(a). The second option is to use one single-task NN like those presented in Section 2.2 (see Figure 2.2(b)). The third option is to consider the P NNs sharing one or more of their hidden layers, and in this way we will have a multi-task NN, see Figure 2.2(c). Generally, a multi-task NN has three main advantages compared to using P different single-task NNs. Firstly, it noticeably improves the training time as we optimize just one NN rather than P different NNs. Secondly, as the countries are likely to share some common behaviours in their mortality evolution, such as the long-term trend of improving mortality, there will likely be mutual benefits for all the P tasks by training them together, see Crawshaw (2020). Thirdly, the multi-task neural network will operate on a single large dataset rather than P smaller datasets, thereby capturing a greater amount of information and leading to more robust predictions.

At this point, we pose a different question: what is the advantage of using a multi-task NN (see Figure 2.2(c)) compared to a single-task NN, as presented in Section 2.2 (see Figure 2.2(b))? The primary advantage is that a multi-task NN not only shares knowledge across related tasks through shared layers (as in single-task NNs) but also enables task-specific specialization via country-specific layers. For instance, when a single-task NN is trained on a large set of countries, it can become dominated by the majority countries—those with similar mortality trends—while minority countries, such as the US and Japan, which exhibit distinct mortality patterns, tend to be under-represented. This imbalance often leads to poorer predictions for the minority countries. In Section 2.5, among other things, we will evaluate whether the multi-task structure in 2.2(c), with its country-specific layers designed to capture unique mortality patterns, can address this issue effectively.

2.3.1 Architecture of the multi-task NNs for mortality forecasting

Similarly to the NNs discussed in Section 2.2, the multi-task NNs will be of the feed-forward type. They will have P input layers, one for each country, where the variables are calendar year, age, country, and gender. There are then P different embedding lay-



(a) Single-task NN

(b) Single-task NN for multi-population

(c) Multi-task NN

Figure 2.2: Illustrations of single and multi-task NNs for mortality prediction.

ers where each categorical variable, i.e. age, country and gender, is transformed into a vector belonging to $\mathbb{R}^5$ as explained in Section 2.2.2. These embedding layers are fully connected to two hidden layers with 128 neurons and tanh activation function, following Richman (2022). The second of these intermediate layers is then fully connected to a third hidden layer with 64 neurons and tanh activation function. From the third hidden layer, there are ramifications with P country-specific hidden layers having 32 neurons and tanh activation function. Finally, these P layers are connected to P output layers where the activation function is of Sigmoid type. Figure 2.3 reports a graphical representation of the just-described NN, which will be referred to as MT1 in the remainder of this paper.

In more formal terms, the layers of MT1 will be computed as follows:

- Input layer:
$$\tilde{\mathbf{X}}_p = (t, x, g, p), \quad p = 1, \dots, P. \tag{2.7}$$
- Embedding layer:
$$\mathbf{X}_p = (t, \mathbf{x}, \mathbf{g}, \mathbf{p}) \in \mathbb{R}^{16}, \quad p = 1, \dots, P, \tag{2.8}$$
$$\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_P) \in \mathbb{R}^{16 \cdot P}. \tag{2.9}$$
- Hidden layer 1:
$$\mathbf{Z}^{(1)} = f^{(1)}(\mathbf{c}^{(1)} + \mathbf{B}^{(1)}\mathbf{X}) \in \mathbb{R}^{128}, \tag{2.10}$$
where $\mathbf{c}^{(1)} \in \mathbb{R}^{128}$, $\mathbf{B}^{(1)} \in \mathbb{R}^{128 \times (16 \cdot P)}$, and $f^{(1)} = \tanh$.
- Hidden layer 2:
$$\mathbf{Z}^{(2)} = f^{(2)}(\mathbf{c}^{(2)} + \mathbf{B}^{(2)}\mathbf{Z}^{(1)}) \in \mathbb{R}^{128}, \tag{2.11}$$
where $\mathbf{c}^{(2)} \in \mathbb{R}^{128}$, $\mathbf{B}^{(2)} \in \mathbb{R}^{128 \times 128}$, and $f^{(2)} = \tanh$.
- Hidden layer 3:
$$\mathbf{Z}^{(3)} = f^{(3)}(\mathbf{c}^{(3)} + \mathbf{B}^{(3)}\mathbf{Z}^{(2)}) \in \mathbb{R}^{64}, \tag{2.12}$$
where $\mathbf{c}^{(3)} \in \mathbb{R}^{64}$, $\mathbf{B}^{(3)} \in \mathbb{R}^{64 \times 128}$, and $f^{(3)} = \tanh$.

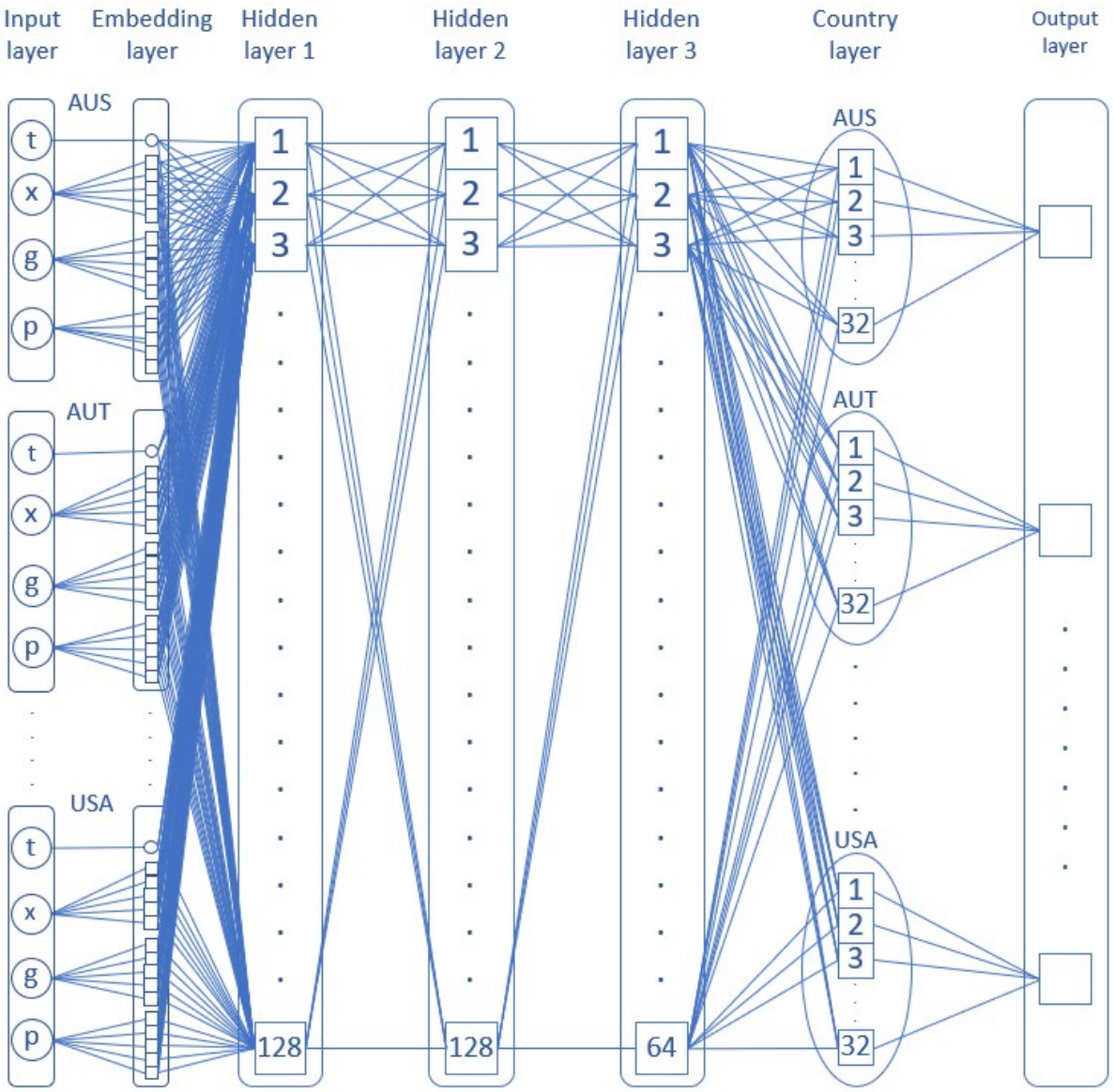


Figure 2.3: Graphical representation of the multi-task NN MT1.

- Country specific layers:

$$\mathbf{Z}_p^{(4)} = f^{(4)}(\mathbf{c}_p^{(4)} + \mathbf{B}_p^{(4)}\mathbf{Z}^{(3)}) \in \mathbb{R}^{32}, \quad p = 1, \ldots, P, \tag{2.13}$$

where $\mathbf{c}_p^{(4)} \in \mathbb{R}^{32}$, $\mathbf{B}_p^{(4)} \in \mathbb{R}^{32\times 64}$, and $f^{(4)} = \tanh$.

- Output layers:

$$Z_p^{(5)} = f^{(5)}(\mathbf{c}_p^{(5)} + \mathbf{B}_p^{(5)}\mathbf{Z}_p^{(4)}) \in \mathbb{R}, \quad p = 1, \ldots, P, \tag{2.14}$$

where $\mathbf{c}_p^{(5)} \in \mathbb{R}$, $\mathbf{B}_p^{(5)} \in \mathbb{R}^{32}$, and $f^{(5)} = \text{sigmoid}$.

2.3.2 Clustering of the third hidden layer

When considering a group of countries, it is natural that some share similar mortality trends while differing from others. These similarities and differences can stem from various social, economic, and geographical factors. For example, the Scandinavian countries, characterised by high wealth levels, extensive social welfare, and geographical proximity, will likely exhibit similar mortality evolutions. To enhance the performance of the multi-task network, we propose clustering the third hidden layer. This approach allows clusters of countries with similar mortality trends to share additional parameters, creating a hierarchical network structure. In this design, lower layers (i.e., hidden layers 1 and 2) capture the overall mortality trend across all countries, while the higher layer (i.e., hidden layer 3) extracts patterns shared by clusters of countries with similar trends. Finally, each country-specific layer learns the distinct mortality pattern for its respective country. To identify countries with similar survival patterns effectively, we can analyze historical mortality data using specific techniques that group countries into homogeneous sets known as clusters. For relevant studies on clustering techniques in the context of mortality forecasting, see Danesi et al. (2015), Nandini and Sanjjushri (2023), and Carracedo et al. (2018).

Having regard to the above discussion, we aim to assess the advantages of clustering the P countries based on their past mortality experiences and construct a new NN architecture that incorporates this clustering. To achieve this, we implement a two-step procedure for each $K = 2, 3$, where K denotes the number of clusters:

1. We use K-means clustering for grouping the P countries into K groups, see Scitovski et al. (2021). In order to do that, we consider the observed changes, in a chosen training period, of the following metrics:

   - Life expectancy for a newborn, truncated at age 90, see Dickson et al. (2019):

$$\mathring{e}_{0:\overline{90}|,t} = \sum_{x=1}^{90} {}_{x-1}p_{0,t}\Big(1 - \frac{1}{2}q_{0+x-1,t}\Big), \tag{2.15}$$

   where $q_{x,t}$ and ${}_hp_{x,t}$ are respectively the 1 year probability of death at age x in year t and the probability of surviving for h years for an individual aged x in year t. These quantities can be derived from the mortality rates $m_{x,t}$ using the following formulas:

$$q_{x,t} = \frac{m_{x,t}}{1 + \frac{1}{2}m_{x,t}}, \tag{2.16}$$

$$ {}_hp_{x,t} = \prod_{j=1}^{h}(1 - q_{x+j-1,t}). \tag{2.17} $$

- Standard deviation of the lifetime of a newborn, truncated at age 90:

$$ \mathrm{SD}_{0:\overline{90|},t} = \sqrt{\sum_{x=0}^{89} {}_{x|1}q_{0,t}\ (x - \mathring{e}_{0:\overline{90|},t})^2 + {}_{90}p_{0,t}(90 - \mathring{e}_{0:\overline{90|},t})^2}, \tag{2.18} $$

  where ${}_{h|1}q_{x,t}$ represents the deferred 1 year probability of death between ages $x+h$ and $x+h+1$ for an individual of age x in year t, and is given by

$$ {}_{h|1}q_{x,t} = {}_hp_{x,t}q_{x+h,t}. \tag{2.19} $$

2. For each K, we build the NN MTK, similar to MT1 but with K clustered hidden layers instead of hidden layer 3. These clustered hidden layers have 64 neurons and a tanh activation function, and are fully connected to hidden layer 2. Furthermore, they are connected with the country-specific layers based on the following rule: if a country is in cluster k, with $k = 1, \ldots, K$, then its country-specific layer is fully connected with cluster layer k. For the remaining parts of the NN, i.e. input layers, embedding layers, hidden layer 1, hidden layer 2, country-specific hidden layers, and output layers, they are specified as in MT1. Formulas for calculating input layer, embedding layer and hidden layers 1 and 2 are the same of (2.7)-(2.11). For the cluster layers, we have

$$ \mathbf{Z}_k^{(3)} = f^{(3)}(\mathbf{c}_k^{(3)} + \mathbf{B}_k^{(3)}\mathbf{Z}^{(2)}) \in \mathbb{R}^{64}, \quad k = 1, \ldots, K, \tag{2.20} $$

   where $\mathbf{c}_k^{(3)} \in \mathbb{R}^{64}$, $\mathbf{B}_k^{(3)} \in \mathbb{R}^{64\times 128}$, and $f^{(3)} = \tanh$. For the country specific layers, we have

$$ \mathbf{Z}_p^{(4)} = f^{(4)}(\mathbf{c}_p^{(4)} + \sum_{k=1}^{K} I_{p,k}\mathbf{B}_p^{(4)}\mathbf{Z}_k^{(3)}) \in \mathbb{R}^{32}, \quad p = 1, \ldots, P, \tag{2.21} $$

   where $\mathbf{c}_p^{(4)} \in \mathbb{R}^{32}$, $\mathbf{B}_p^{(4)} \in \mathbb{R}^{32\times 64}$, $f^{(4)} = \tanh$, and $I_{p,k} = 1$ if country p belongs to cluster k and $I_{p,k} = 0$ otherwise. Finally, the formula for the output layer is the same as (2.14).

The architectures of MT2 and MT3 can be found respectively in Figures 2.4 and 2.5.

Regarding the choice to implement NNs with only two or three clusters, this decision was based on preliminary experiments using a larger number of clusters (up to five). These experiments did not reveal any significant improvement in out-of-sample performance, while increasing both the number of parameters and, consequently, the computational time required for training the NNs.

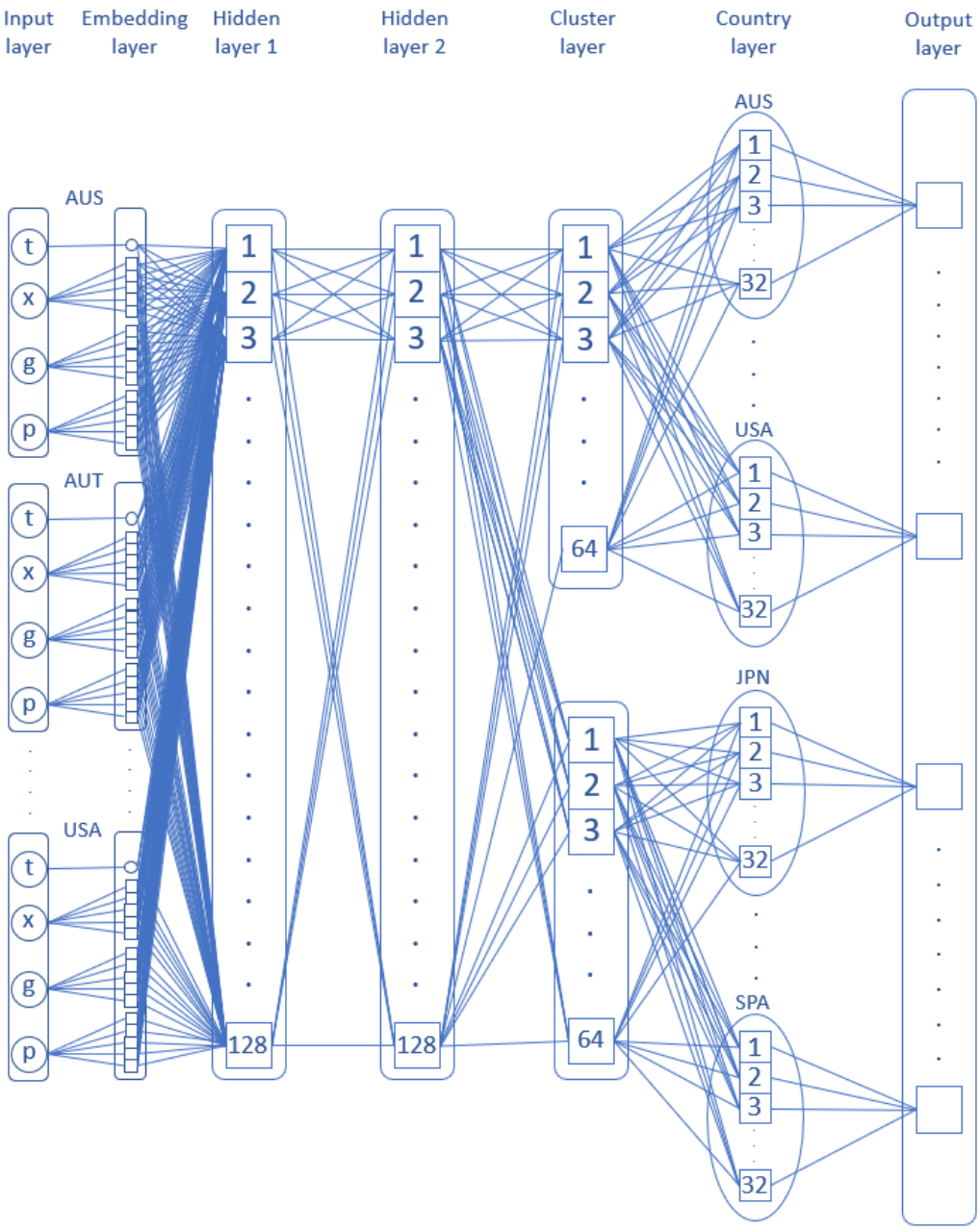


Figure 2.4: Graphical representation of the multi-task NN MT2.

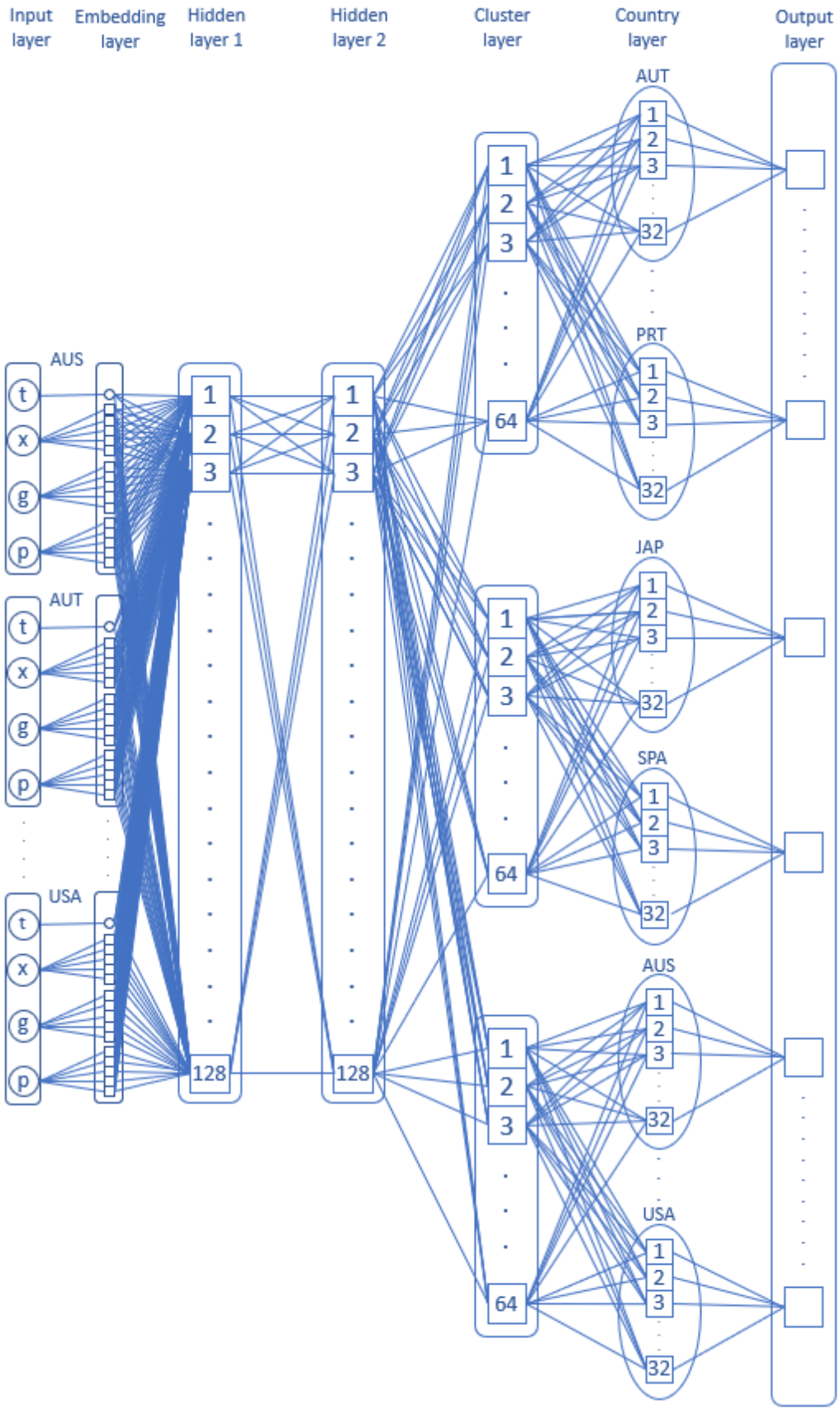


Figure 2.5: Graphical representation of the multi-task NN MT3.

2.4 Data, clustering and training

The choice of the countries we consider in the quantitative analysis is based on three factors: firstly, they must have data available in the HMD[1]. Secondly, historical data series for these countries must be complete from the year 1950 onwards. Thirdly, each selected country must have had a population of at least 3 million in 1950. In light of this, we consider historical mortality data for males and females from $P = 17$ countries: Australia, Austria, Belgium, Canada, Denmark, England & Wales, Finland, France, Italy, Japan, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the US. For these countries, we consider the yearly central mortality rates obtained from HMD in three different age bands: 0-89, 20-89, and 55-89 to test the sensitivity of the different approaches with respect to the age band. Regarding the choice of the time interval, we considered, following Richman and Wüthrich (2021), a 50-years training period (1950-1999) and a 20-years test period (2000-2019). Finally, in order to study how the performance of the models varies based on the length of the training period, we also considered the following training sets: 1955-1999, 1960-1999, 1965-1999, 1970-1999, 1975-1999, and 1980-1999 (using 20-89 as reference age range).
The results of clustering using the approach described in Subsection 2.3.2 are reported in Table 2.2. Looking at the composition of the resulting clusters, we observe that Japan and Spain are grouped together in both cases. In MT2, Portugal joins these two countries, whereas in MT3 it clusters instead with several European nations, namely Austria, Belgium, Finland, France, and Italy. Conversely, Australia, Canada, Denmark, England & Wales, the Netherlands, Norway, Sweden, Switzerland, and the United States consistently fall into the same cluster across both cases.
Regarding the training of the multi-task NNs, we used the following hyperparameters: 150 epochs when using 55-89 age band, and 250 epochs when using 20-89 and 0-89 age bands, batch size equal to 32, learning rate equal to 0.0005, Adam optimizer, and mean squared error as loss function:

$$L = \sum_{p=1}^{P} \frac{1}{n_p} \sum_{j=1}^{n_p} w_j^{(p)} (m_j^{(p)} - \hat{m}_j^{(p)})^2 \tag{2.22}$$

where n_p is the total number of observations for country p, and $w_j^{(p)}$ is the relative weight. We set $w_j^{(p)} = 1$ in the unweighted case and $w_j^{(p)} = \frac{1}{m_j^{(p)}}$ in the weighted case. In the following, results obtained using multi-task NNs are denoted by MT1, MT2, and MT3, in the unweighted case, and by MT1 w, MT2 w, and MT3 w, in the weighted case. Finally, we repeated the training of each NN 10 times in order to ensure robustness towards the effects of randomness in the training process.

2.5 Results

In this section, we compare goodness of the forecasts obtained using multi-task NNs, MT1, MT2 and MT3 (both in the weighted and unweighed case), the single-task NNs DEEP1, DEEP2, DEEP3, DEEP4, DEEP5 and DEEP6, and 3 widely-used stochastic mortality models from the literature - the single population version of the LC model, see

[1]Human Mortality Database: www.mortality.org.

Table 2.2: Results of clustering.

MT1		MT2		MT3	
Country	**Cluster**	**Country**	**Cluster**	**Country**	**Cluster**
Australia		Japan		Australia	
Austria		Portugal	1	Canada	
Belgium		Spain		Denmark	
Canada		Australia		England & Wales	
Denmark		Austria		Netherlands	1
England & Wales		Belgium		Norway	
Finland		Canada		Sweden	
France		Denmark		Switzerland	
Italy	1	England & Wales		USA	
Japan		Finland	2	Japan	2
Netherlands		France		Spain	
Norway		Italy		Austria	
Portugal		Netherlands		Belgium	
Spain		Norway		Finland	3
Sweden		Sweden		France	
Switzerland		Switzerland		Italy	
USA		USA		Portugal	

Lee and Carter (1992),

$$\ln(m_{x,t}^{(g,p)}) = \alpha_x^{(g,p)} + \beta_x^{(g,p)}\kappa_t^{(g,p)}, \tag{2.23}$$

the single population version of the CBD model[2], see Cairns et al. (2006),

$$\ln(m_{x,t}^{(g,p)}) = \kappa_t^{(1,g,p)} + \kappa_t^{(2,g,p)}(x - \bar{x}), \tag{2.24}$$

and a version of the ACF model, see Chen and Millossovich (2018), used for modelling simultaneously both gender and a set of different countries,

$$\ln(m_{x,t}^{(g,p)}) = \alpha_x^{(g,p)} + B_x K_t + \beta_x^{(g)} k_t^{(g)} + \beta_x^{(g,p)} k_t^{(g,p)}. \tag{2.25}$$

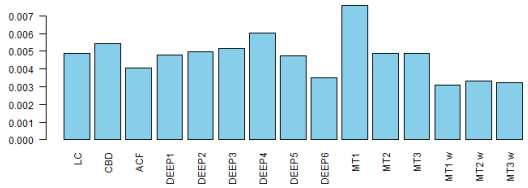
Here $\alpha_x^{(g,p)}$, $\beta_x^{(g)}$, $\beta_x^{(g,p)}$, and B_x are age-dependent parameters, $\bar{x} = 72$, the average over the population age range 55-89, while $k_t^{(g)}$, $k_t^{(g,p)}$, $k_t^{(1,g,p)}$, $k_t^{(2,g,p)}$ and K_t are time-dependent stochastic factors. Here $k_t^{(g,p)}$ (in the LC model), $k_t^{(1,g,p)}$, $k_t^{(2,g,p)}$ and K_t are modelled as a random walk with drift, while $k_t^{(g)}$ and $k_t^{(g,p)}$ (in the ACF model) are modelled as an $AR(1)$. Notice that unlike the ACF model, the LC and CBD models treat different countries independently. Both the fitting and the forecasting of these three models are obtained using the StMoMo package, see Villegas et al. (2018).

The comparison of models results is based on three metrics: mean absolute forecasting error (MAFE) for individual yearly death rates, for life expectancy and for standard deviation of the lifetime. When calculating these metrics, both male and female populations are considered. Figures 2.6, 2.7, and 2.8 report the three metrics respectively for the 55-89, 20-89, and 0-89 age ranges while using 1950-1999 as training period and 2000-2019 as test period. Figures 2.10-2.18 extend this analysis by showing the three metrics for individual countries. Figure 2.9 shows the evolution of the three metrics using different lengths for the training period using 20-89 as age range. Finally, Table 2.3 summarises the total number of parameters in each approach for the three age ranges considered here.[3]

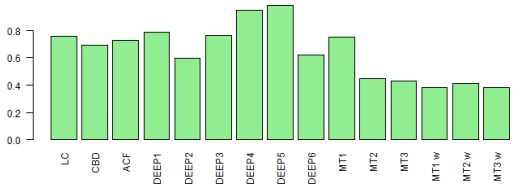
[2]CBD model is only tested in the 55-89 age range, for which it was designed.

[3]Notice that MT1 w, MT2 w, and MT3 w have the same number of parameters of respectively MT1, MT2, and MT3.

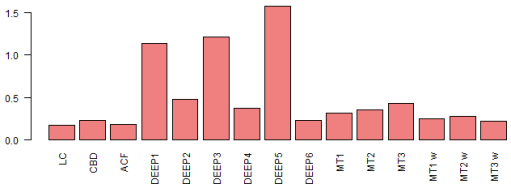
We observe that there is variability in the different approaches performance based on the metric and age range considered. In Figure 2.6, we notice how multi-task NNs show noticeable results in the 55-89 age range both with and without the weighting scheme. Indeed, with the sole exception of MT1, they outperform all other approaches in terms of life expectancy MAFE. Multi-task NNs with weighting scheme results appear to be the best ones for mortality rates MAFE, while in standard deviation MAFE they are outperformed only by Lee-Carter and ACF models. Focusing on 20-89 age range in Fig-



(a) MAFE - Mortality Rates



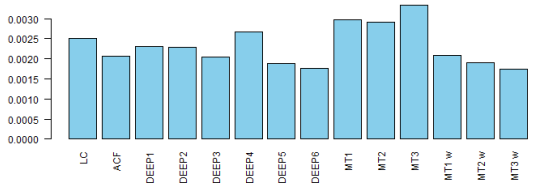
(b) MAFE - Life Expectancy



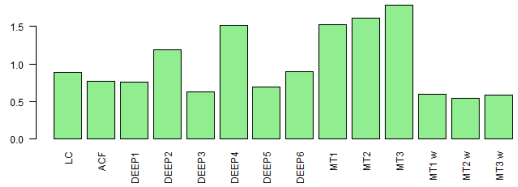
(c) MAFE - Standard Deviation

Figure 2.6: Comparison of MAFE for Mortality Rates, Life Expectancy, and Standard Deviation. Age range: 55-89.

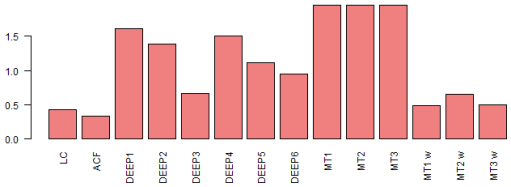
ure 2.7, we notice a big difference with respect to the 55-89 age range case. Firstly, multi-task NNs without a weighting scheme turn out to be the worst ones for all three the metrics considered here. In contrast, multi-task NNs with a weighting scheme still show good results. Indeed, they are among the best ones for mortality rate MAFE, alongside DEEP5 and DEEP6, for life expectancy MAFE, alongside DEEP3, and for standard deviation, where they are outperformed only by Lee-Carter and ACF models. The reason for such a big difference between the performance of multi-task NNs with and without weighting scheme, especially for life expectancy and standard deviation, is likely due to the training of the NNs. Indeed, mortality rates at lower ages are underestimated, due to their lower magnitude, during the training process which tends to place more emphasis on observations with higher magnitude, such as the ones at older ages. As a consequence, the NNs will produce poor forecast for lower ages mortality rates and, as a consequence, bigger errors for life expectancy and standard deviation which are heavily influenced by mortality at early ages. Finally, focusing on Figure 2.8, we observe a similar pattern also when considering the 0-89 age range. Indeed, the multi-task NNs trained without a weighting scheme result in the poorest performance for all the three metrics considered. Also multi-task NNs with a weighting scheme have a weaker performance compared to the other two cases considered previously. In fact, we notice how Lee-Carter and ACF models, and DEEP1, DEEP3, and DEEP5 single-task NNs have better performance compared to them. Nevertheless, the advantage of using a weighting



(a) MAFE - Mortality Rates



(b) MAFE - Life Expectancy

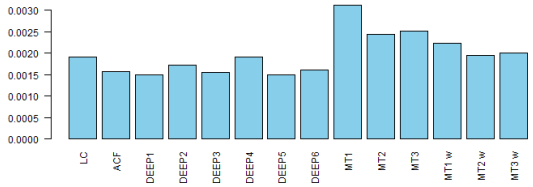


(c) MAFE - Standard Deviation

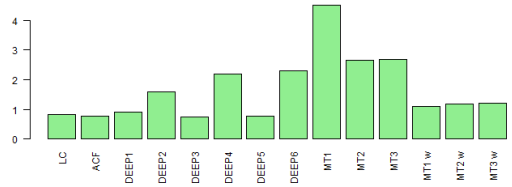
Figure 2.7: Comparison of MAFE for Mortality Rates, Life Expectancy, and Standard Deviation. Age range: 20-89.

scheme is still important for multi-task NNs, especially when considering life expectancy and standard deviation. Finally, here we can notice the benefit of clustering the third hidden layer in multi-task NNs (notice that the clustering is based on historical values of life expectancy and standard deviation in the age range 0-89). Indeed, both MT2 and MT3 show an improvement to MT1 in all the metrics considered here. We also observe that when introducing a weighting scheme in the training of the NNs this benefit tends to disappear. In Figures 2.10-2.18, the MAFE by country and approach is shown. Specifically, the height of vertical black line represents the global MAFE (i.e. the same metric showed in Figures 2.6-2.8), while the coloured dots represent the MAFE for individual countries. Focusing on the US and Japan, i.e. two countries with particular pattern in the evolution of mortality in the last decades, we notice how the multi-task NNs with a weighting scheme provide generally better results with respect to single-task NNs if we consider the 55-89 age range. When widening the age range to 20-89 and 0-89, the advantage of multi-task NNs on the two countries tend to disappear, with their MAFE being often above the one of single-task NNs. Overall when considering all countries, we notice how in the 55-89 age range case, multi-task NNs tend to have lower dispersion across the countries' MAFEs compared to single-task NNs.

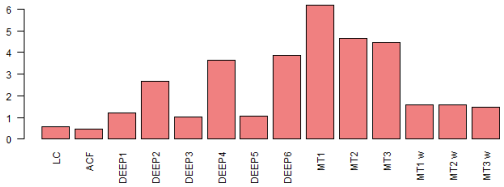
In Figure 2.9, the minimum MAFE by approach, training period (while keeping fixed the age range 20-89), and metric is reported. The minimum MAFEs for single-task NNs, multi-task NNs without weighting scheme, multi-task NNs with weighting scheme, and stochastic models are obtained as the minimum among DEEP1-DEEP6, MT1-MT3, MT1 w-MT3 w, and LC-ACF, respectively. We can observe how for mortality rates MAFE, single-task NNs generally give the best results with the exception of three training periods: in 1970-1999 and 1955-1999, they are outperformed by multi-task NNs with a weighting scheme, and in 1975-1999, they are outperformed by stochastic models. Although the gap with stochastic models narrows when considering shorter training periods, multi-task NNs performance appears to be less steady, and results to be the best one only in one case. Nevertheless, the benefit of using a weighting scheme is noticeable in all the



(a) MAFE - Mortality Rates



(b) MAFE - Life Expectancy



(c) MAFE - Standard Deviation

Figure 2.8: Comparison of MAFE metrics for Mortality Rates, Life Expectancy, and Standard Deviation. Age range: 0-89.

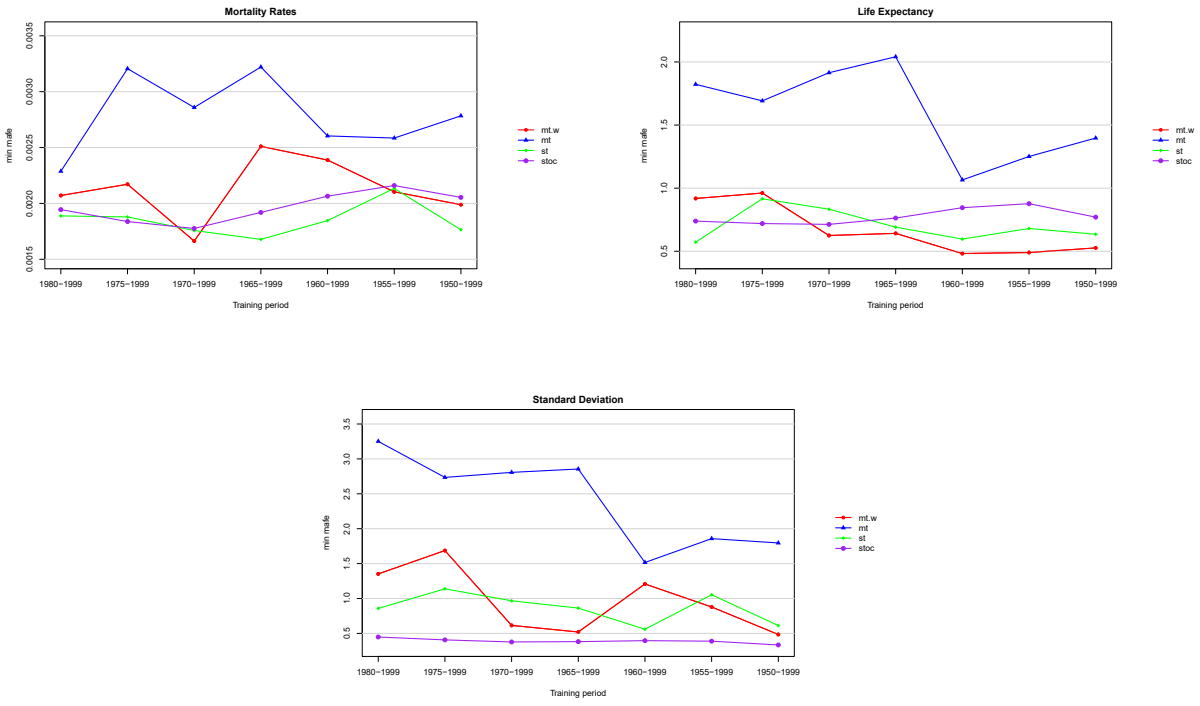


Figure 2.9: Minimum MAFE for single-task NNs, multi-task NNs, and stochastic models by training period and metric considered (age range 20-89).

training periods considered here. Life expectancy is the metric on which the multi-task neural networks perform relatively better. In fact, they achieve the lowest MAFE in five out of seven training periods, from 1970–1999 to 1950–1999. This suggests that, for life expectancy, multi-task neural networks yield the best results when trained on a broader time interval. By contrast, such a pattern is less clear for single-task neural networks and

stochastic models. Finally, the stochastic models outperform NNs in all cases considered while focusing on lifetime standard deviation. They always have the lowest minimum MAFE, despite the gap with NNs getting more narrow in the longest training period, showing a similar trend also found in the mortality rates MAFE. With regards to the NNs, single-task and multi-task alternate with each other in terms of best performance with a consistent result that using a weighting scheme improves the forecasting accuracy of multi-task NNs.

Table 2.3: Number of parameters and data points by approach and age range.

	0-89	**20-89**	**55-89**
LC model	7,820	6,460	4,080
CBD model	-	-	3,400
ACF model	8,240	6,820	4,335
DEEP1	20,386	20,286	20,111
DEEP2	20,386	20,286	20,111
DEEP3	71,458	71,358	71,183
DEEP4	71,458	71,358	71,183
DEEP5	73,506	73,406	73,231
DEEP6	73,506	73,406	73,231
MT1	108,354	106,654	103,679
MT2	116,866	115,166	112,191
MT3	125,378	123,678	120,703
#data points	153,000	119,000	59,500

Focusing on single-task neural networks, several observations can be made from Figures 2.10-2.18 regarding their performance across different metrics and age ranges. In particular, when comparing DEEP5 and DEEP6 with DEEP3 and DEEP4, it appears that adding a skip connection between the embedding layer and the last hidden layer does not lead to an improvement in overall performance. More relevant, instead, is the choice of activation function in the hidden layers relative to the age range considered. Specifically, for the 55–89 age range, networks employing the tanh activation function perform markedly better; for the 0–89 age range, networks using ReLU exhibit superior performance; whereas for the 20–89 age range, no strong preference between the two activation functions is observed.
In conclusion, Table 2.3 provides the details on the number of parameters and data points considered, categorized by approach and age range, for the training period 1950-1999. We can notice how the number of parameters for stochastic models, DEEP1, and DEEP2 is definitely lower than the number of data points in all the age ranges considered. The remaining single-task NNs, MT1, and MT2 exceed the number of data points when the age range is 55–89 but stay below it for wider age ranges. Finally, MT3 has a number of parameters lower than the number of data points only when considering the 0-89 age range. These numbers highlight how multi-task networks may become overparameterised when trained on narrower age ranges, where fewer observations are available. This imbalance may increase the risk of overfitting and adversely affect the models' generalisation performance.

2.6 Conclusion

The results show that using a 50-year training period, the performance of multi-task NNs compared to single-task NNs and traditional stochastic models depends on the metric considered and, especially, on the age range. More specifically, the out-of-sample precision

of multi-task NNs is good with a shorter age range but tends to deteriorate when this age range is increased. This is likely due to the underestimation of lower ages mortality rates that happen in the training period. Adding a weighting scheme to multi-task NNs, markedly improves their performance, especially for life expectancy and standard deviation. Finally, it is noticeable that multi-task NNs with clustering based on past life expectancy and standard deviation show better results only when a weighting scheme is not considered.

When testing the models on shorter training periods, we arrive at a similar conclusion with respect to the 1950-1999 training period case. What is worthy to point out is that traditional stochastic models tend to perform relatively better compared to NNs when considering a short training period.

In terms of future research on multi-task NNs, at least five straightforward developments could be considered: firstly, implementing multi-task NNs where the generic task is based on a categorical variable such as age and gender, rather than, or alongside, country. Secondly, using a different machine learning technique, such as one specifically designed for time series, to cluster the countries based on past mortality experience. Thirdly, a penalization could be added to the loss function of the NNs to ensure that female and male mortality do not diverge, or even that the mortality of different populations does not diverge, achieving some degree of coherence. Fourthly, forecasting mortality rates in a different context, such as by cause of death within a single population. Fifthly, further analysis aiming to study the question of explainability of multi-task NNs could be conducted, see Perla et al. (2024).

Appendix - Results by country

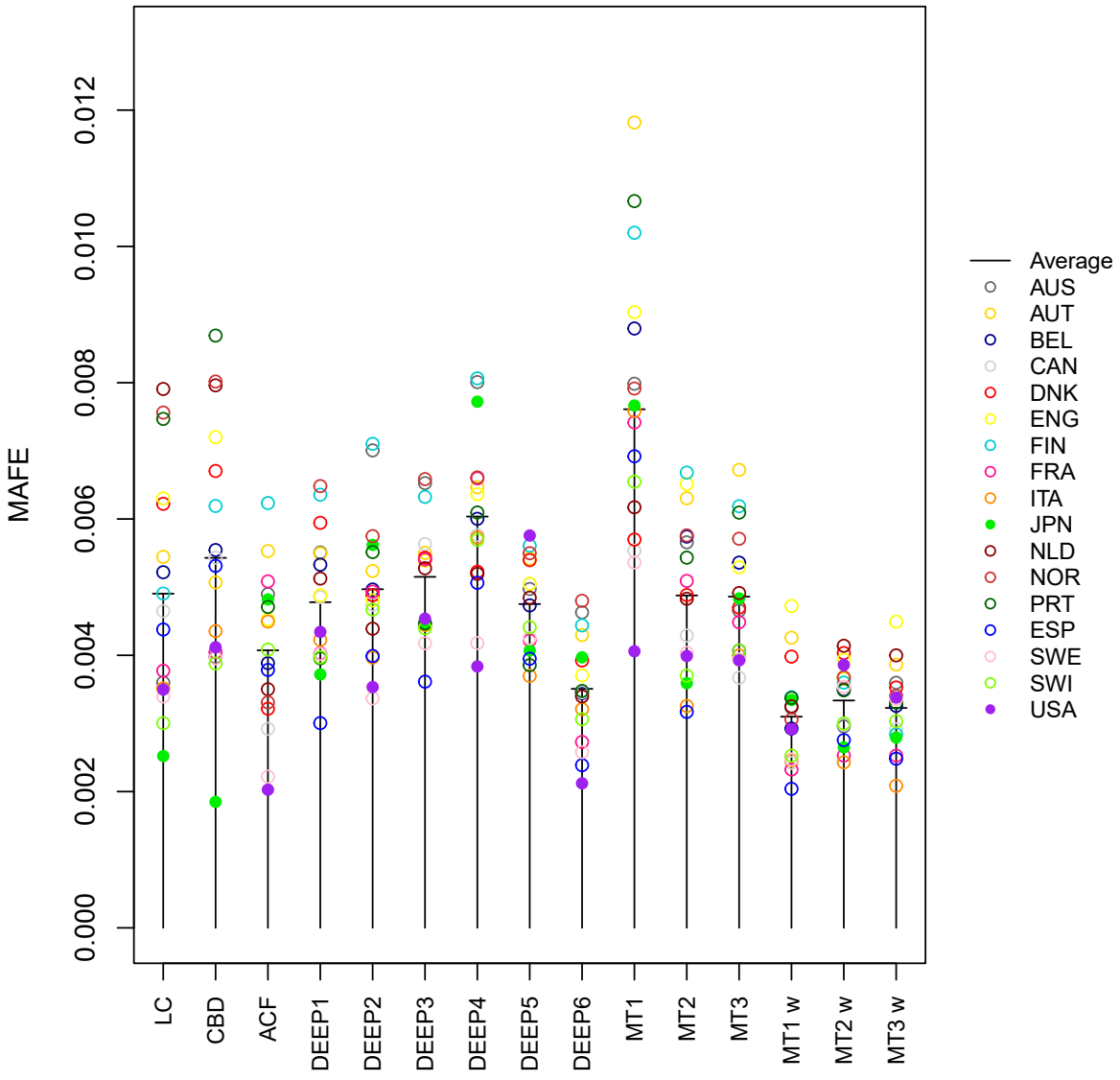


Figure 2.10: Mean absolute forecasting error by country and approach. Metric considered: mortality rate. Age range: 55-89.

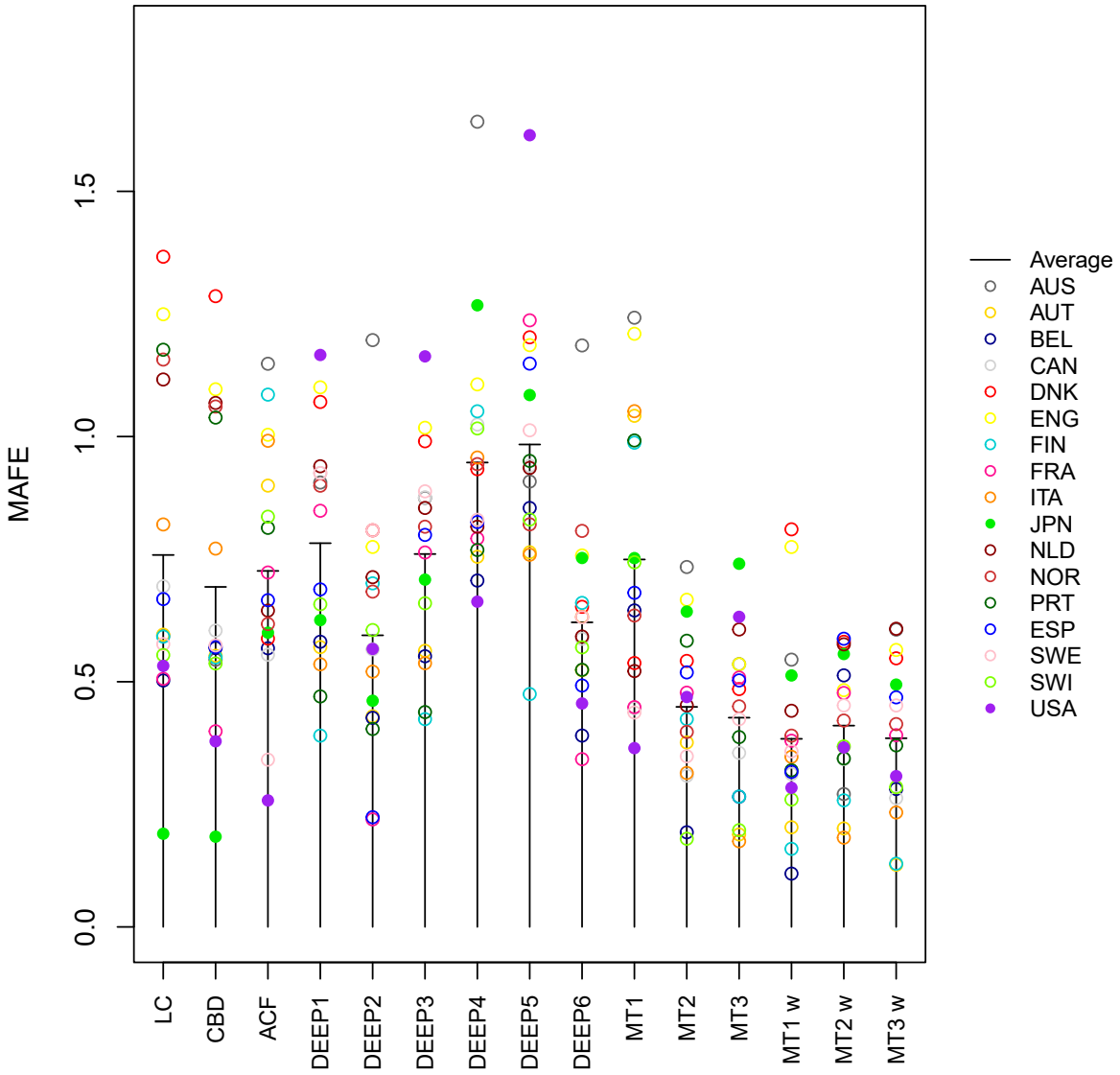


Figure 2.11: Mean absolute forecasting error by country and approach. Metric considered: life expectancy. Age range: 55-89.

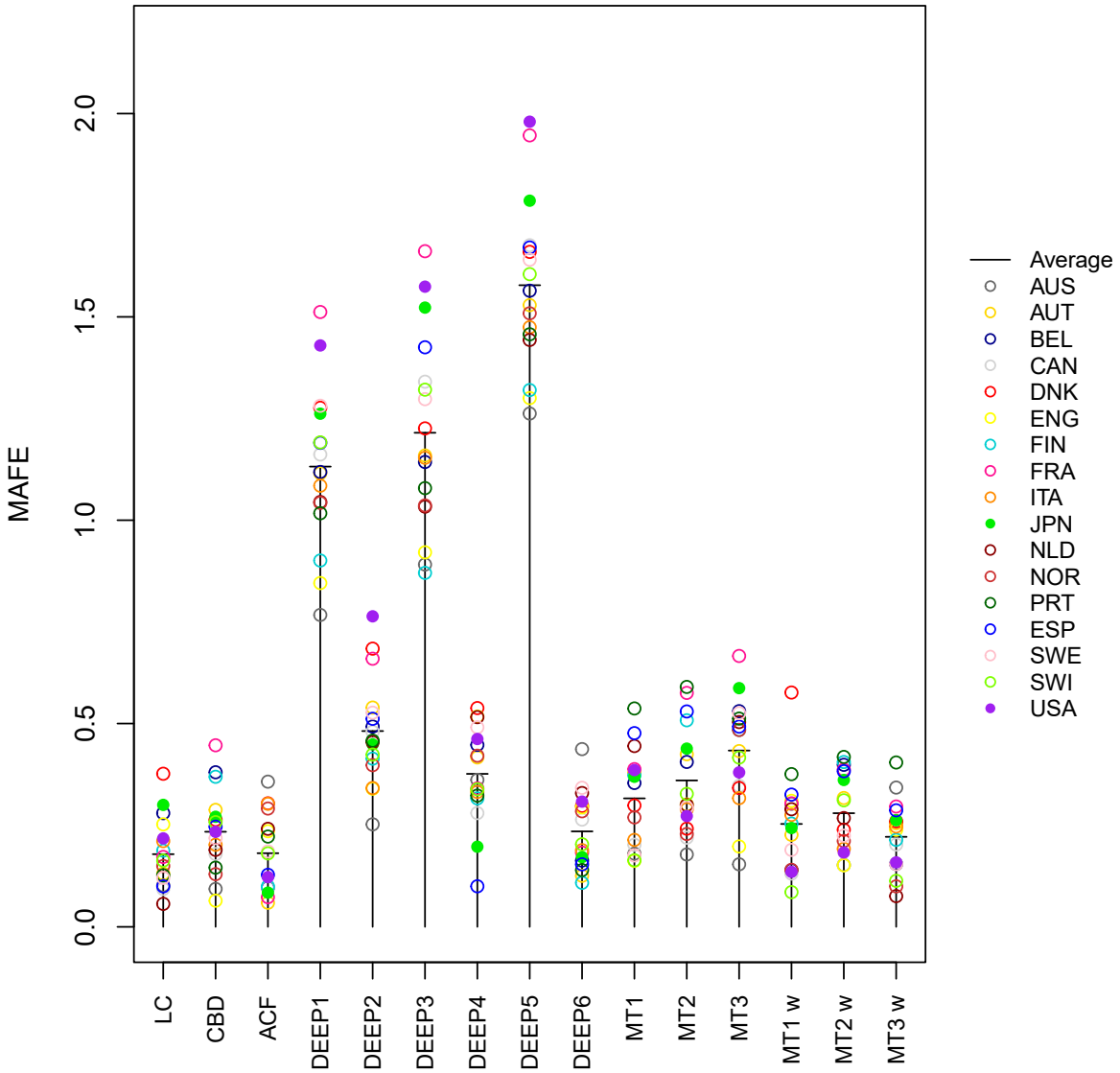


Figure 2.12: Mean absolute forecasting error by country and approach. Metric considered: standard deviation. Age range: 55-89.

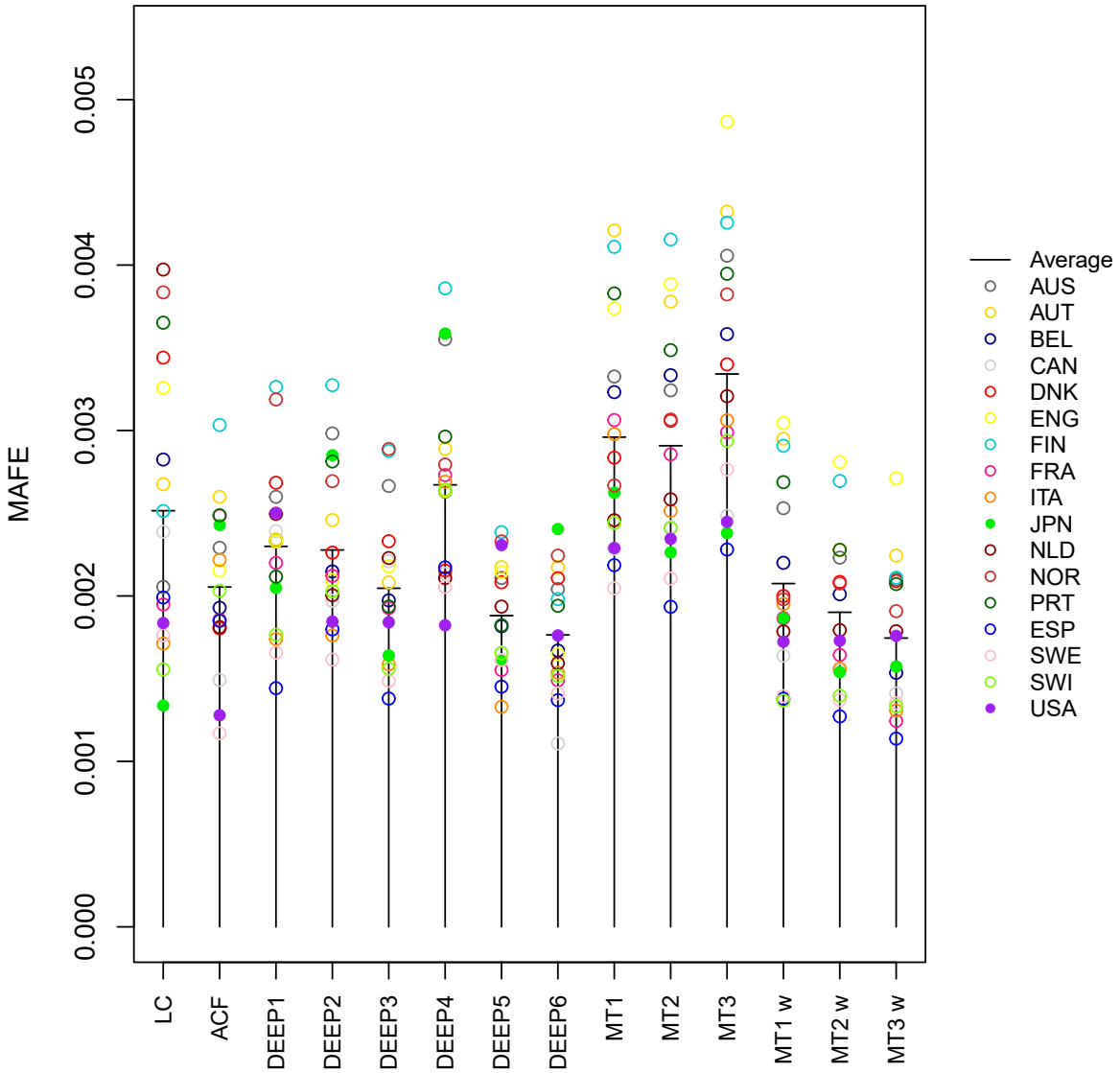


Figure 2.13: Mean absolute forecasting error by country and approach. Metric considered: mortality rate. Age range: 20-89.

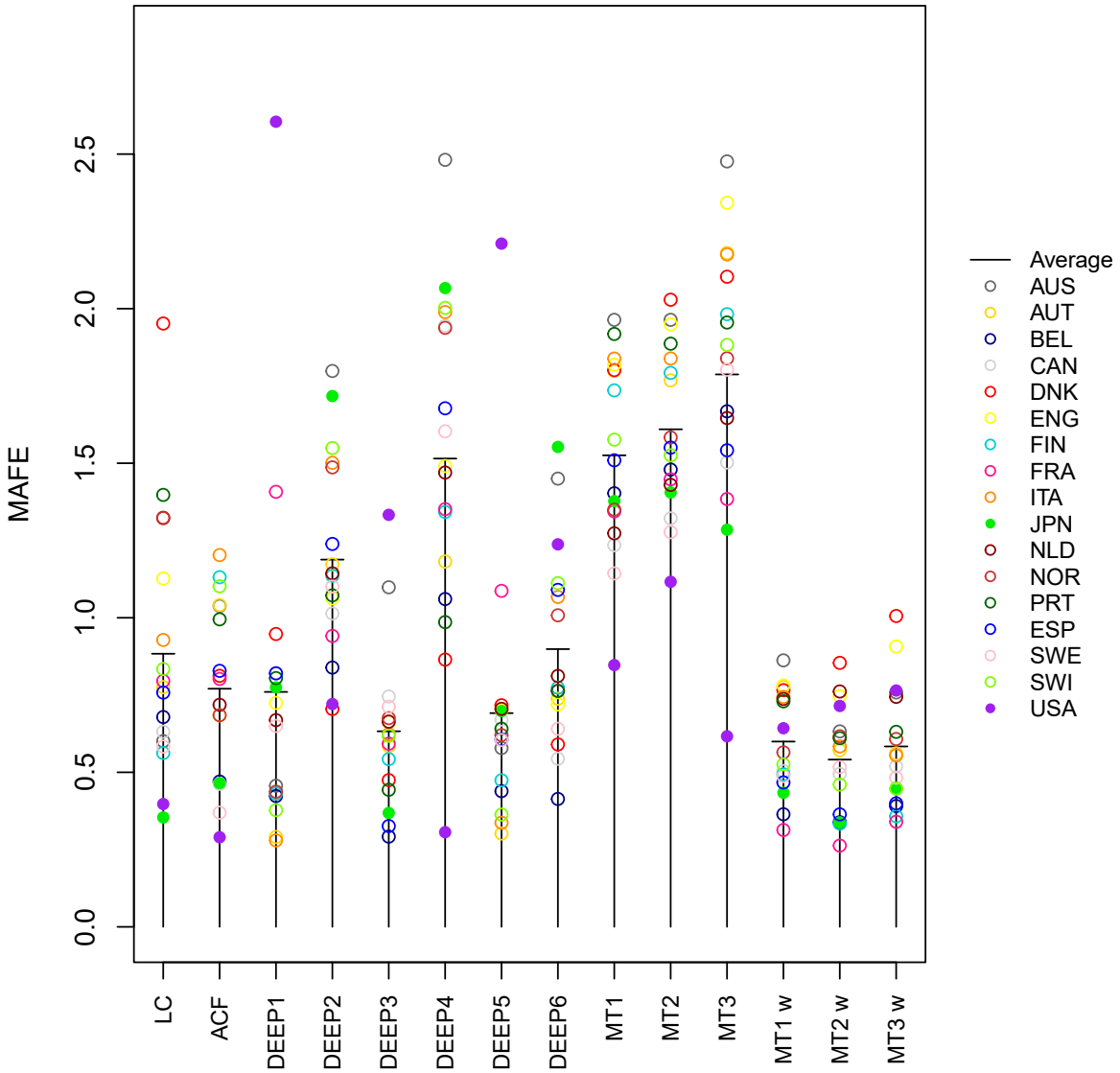


Figure 2.14: Mean absolute forecasting error by country and approach. Metric considered: life expectancy. Age range: 20-89.

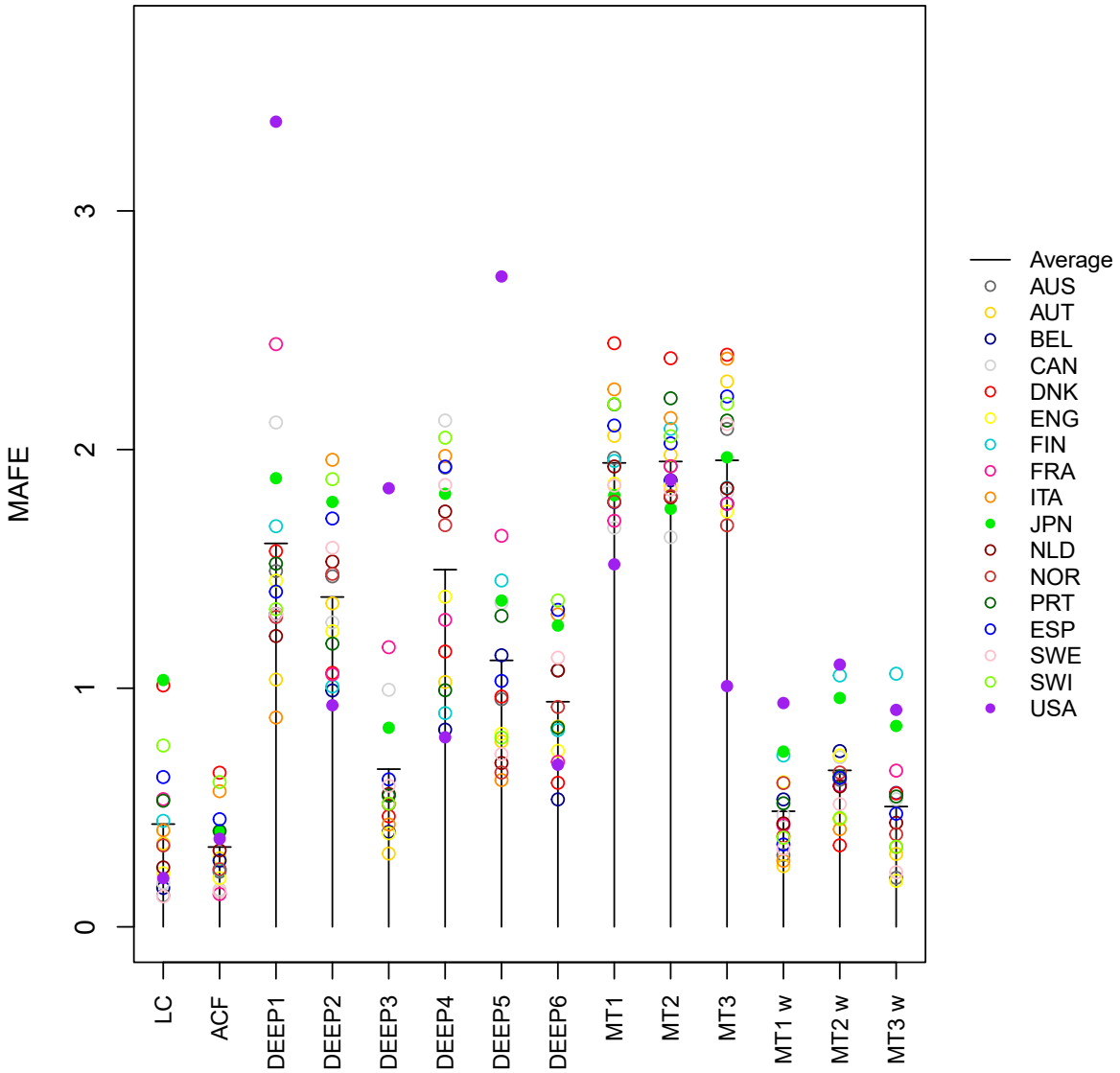


Figure 2.15: Mean absolute forecasting error by country and approach. Metric considered: standard deviation. Age range: 20-89.

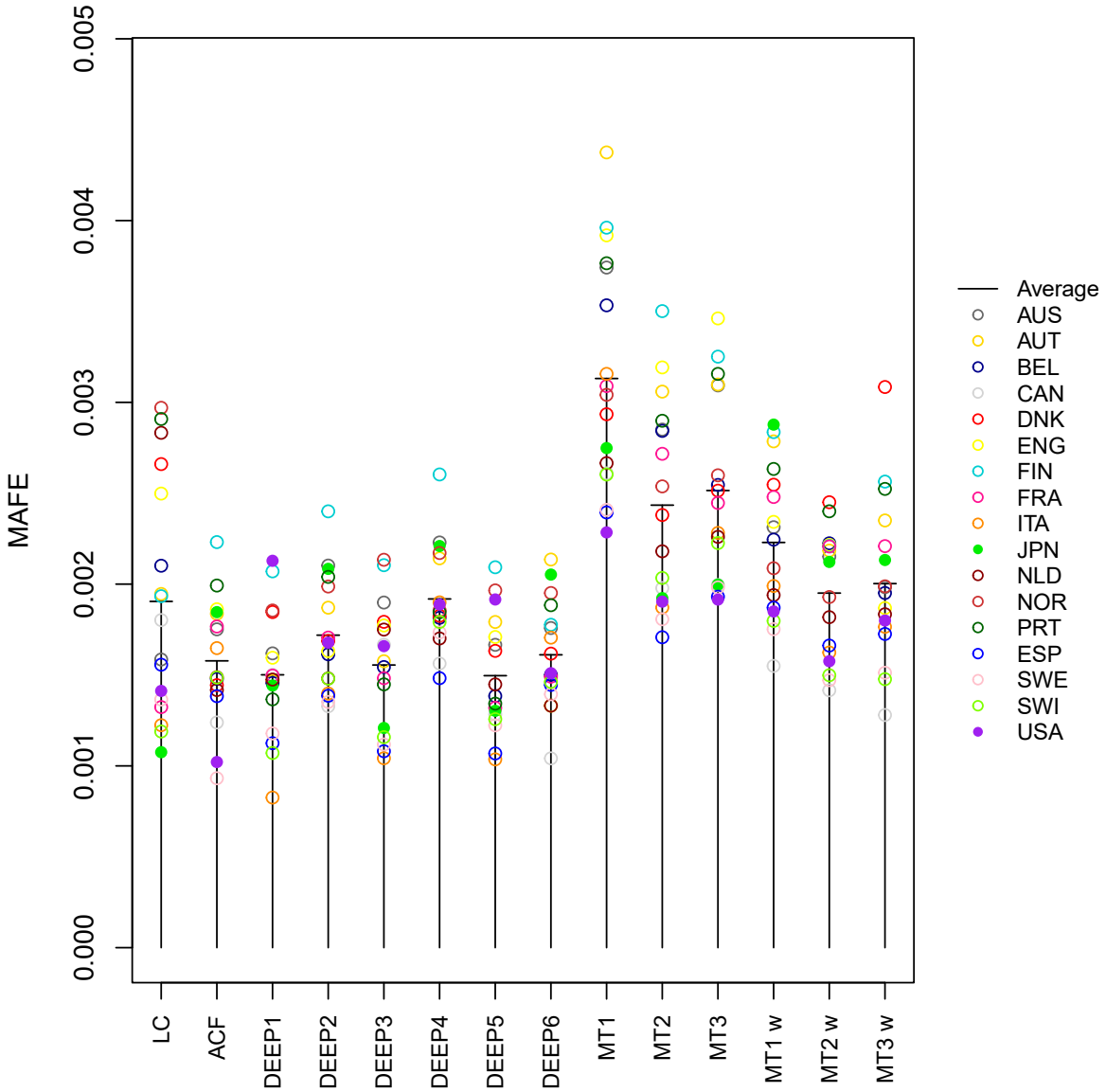


Figure 2.16: Mean absolute forecasting error by country and approach. Metric considered: mortality rate. Age range: 0-89.

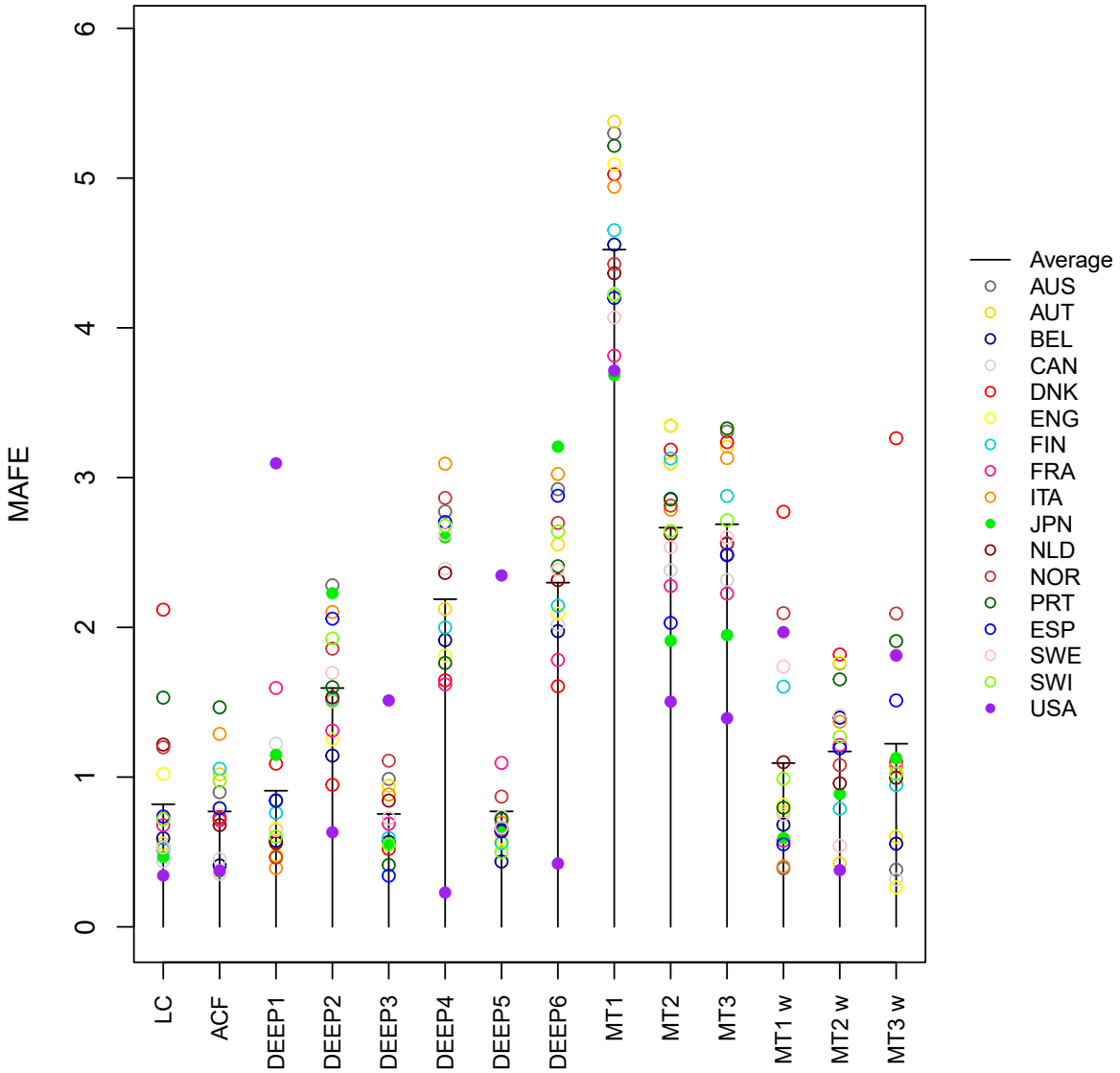


Figure 2.17: Mean absolute forecasting error by country and approach. Metric considered: life expectancy. Age range: 0-89.

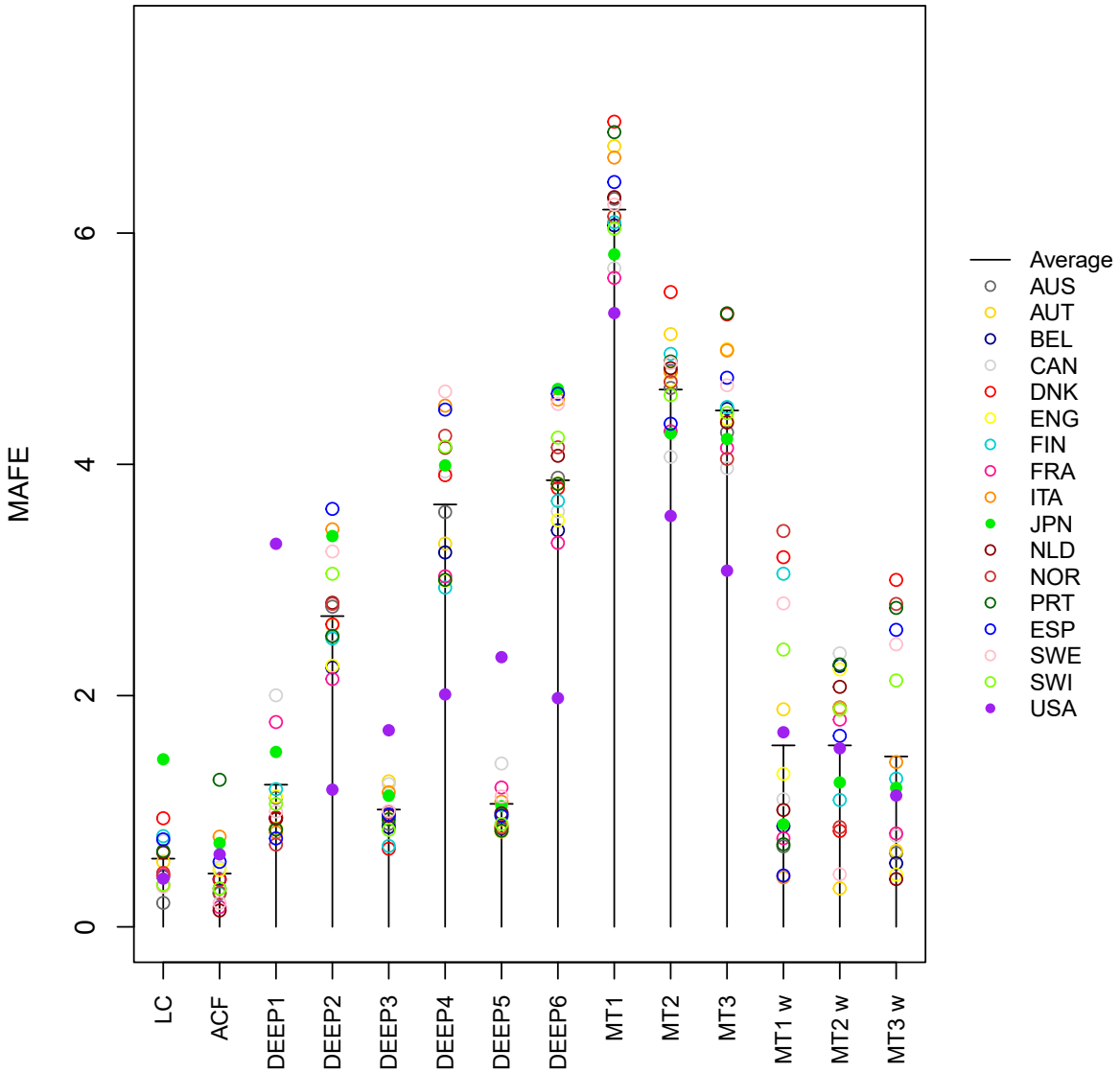


Figure 2.18: Mean absolute forecasting error by country and approach. Metric considered: standard deviation. Age range: 0-89.

Chapter 3

Forecasting mortality rates by cause of death and socio-economic class using neural networks

3.1 Introduction

Mortality trends have changed rapidly over the last century, primarily due to advancements in medicine, technology, and lifestyle. Mortality rates have generally been decreasing in most countries, except for temporary increases caused by wars or pandemics. As a result, it has become essential for governments, policymakers, and private institutions to study these trends in order to understand the main drivers of these changes and to help forecast future mortality more accurately (see Willets et al. (2004), Jones et al. (2020), and Purushotham et al. (2011)).
Alongside the study of the evolution of overall population mortality, increasing importance has been placed on the study of mortality by cause of death (see Villegas et al. (2024)). Traditionally, actuaries have addressed this problem using life tables with multiple decrements (see Schoen (2013)). Decomposing overall mortality by cause of death is crucial for understanding phenomena such as the mortality gap between females and males, which helps in forecasting mortality trends more effectively. For instance, the proportion of deaths caused by infections has decreased significantly over the last century due to the development of vaccines and antibiotics. Figure 3.1 illustrates the historical evolution of the probability of dying at ages 70-79 by cause of death in England between 2001 and 2020.

In conclusion, we argue that forecasting mortality by cause of death is crucial for healthcare planning and policymaking from the perspectives of governments, policymakers, and key industries such as healthcare and financial services. Factors such as ageing populations, emerging diseases, and shifting health trends make this topic increasingly important for resource allocation, improved public health interventions, and enhanced financial planning for insurance and pension systems. Therefore, studying cause-of-death mortality rates is essential not only for forecasting overall mortality rates but also for developing effective public and private policies and interventions.
Among the important papers that have studied this topic with the aim of providing useful tools for forecasting specific cause-of-death mortality rates, we highlight the following: McNown and Rogers (1992), which estimates the parameters of a multiexponential model from historical data and then forecasts them using ARIMA time series; Gaille and Sherris

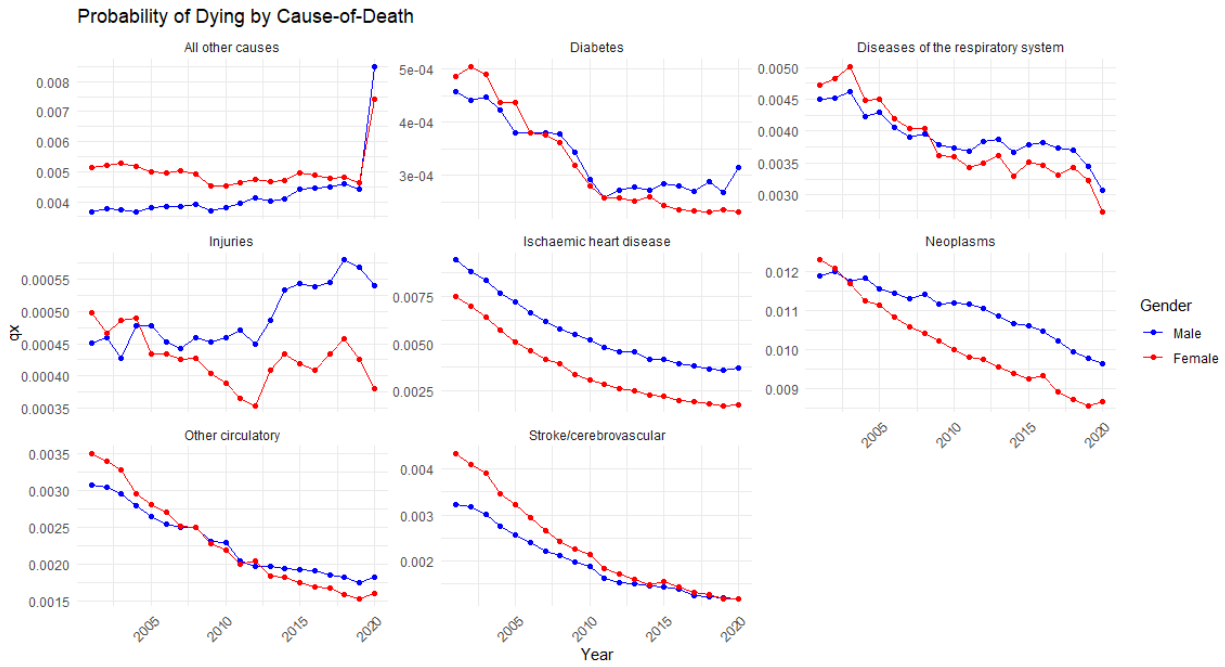


Figure 3.1: Evolution of mortality for males (in blue) and females (in red) by cause of death in England between 2001 and 2020. Age interval 70-79.

(2011), which estimates the Heligman-Pollard mortality model parameters and forecasts them using Vector Error Correction Models; Arnold and Sherris (2013), which uses Vector Error Correction Models to model multiple causes of death simultaneously and capture their time dependencies; Alai et al. (2015), which employs multinomial logistic regression to jointly address different causes of death, using period as one of the independent variables, allowing for the forecasting of mortality metrics without relying on time series; Caselli et al. (2019), which applies the Lee-Carter model and the APC model to forecast mortality rates for each cause of death separately; Li and Lu (2019), which utilises hierarchical Archimedean copulas—a mathematical framework for modelling and analysing the dependence structure among multiple random variables—both to model the relationships among causes of death and to produce forecasts; and finally, Dong et al. (2020) and Zhang et al. (2023), which employ a machine learning technique called tensor decomposition, further discussed in Section 3.2, for multi-population and cause-of-death mortality forecasting.

Alongside the study of mortality by cause of death, we also aim to focus on mortality by socio-economic class and how this factor impacts the dynamics of mortality. Studying mortality by socio-economic class is crucial for several reasons: it helps identify and provide insights into health disparities that have widened over recent decades (see Willets et al. (2004), Clouston et al. (2016), and Mackenbach et al. (2017)), and it informs policymaking aimed at improving public health (see Lu et al. (2014) and Madrigal et al. (2011)). As examples of forecasting mortality for different socio-economic subgroups of the population using stochastic methods, see Villegas and Haberman (2014) and Strozza et al. (2024).

The aim of this chapter is to implement feed-forward neural networks, both of the single-task and multi-task types, for forecasting mortality rates by cause of death, socio-economic class, and their combination. Additionally, the chapter evaluates the forecasting

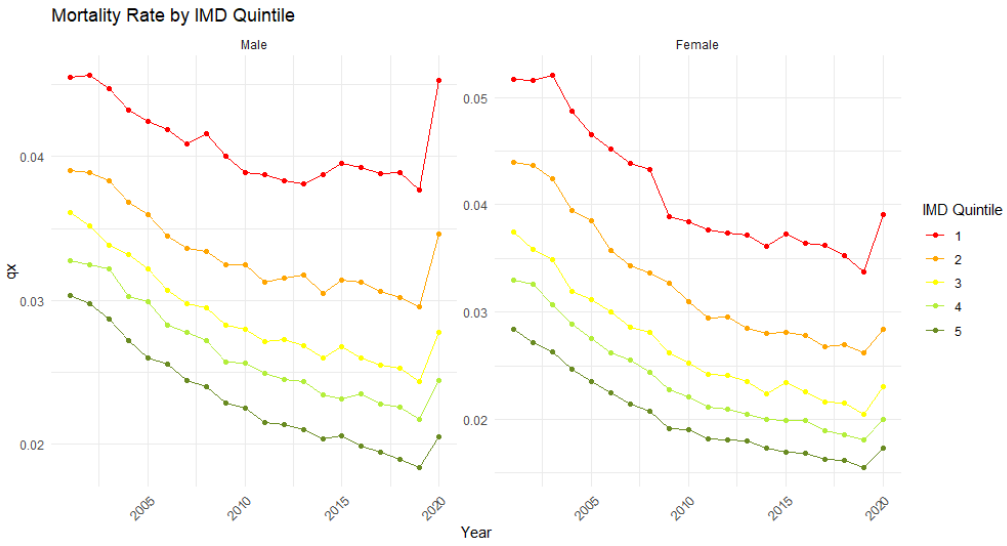


Figure 3.2: Evolution of mortality for males and females by socio-economic class in England between 2001 and 2020. Age interval 70-79. IMD (Index of Multiple Deprivation) 1 corresponds with the most disadvantaged socio-economic class, while IMD 5 with the most affluent.

performance of these neural networks compared to other methodologies existing in the literature: the Lee-Carter model, i.e., a stochastic model which provides forecasts at the single population level, and Penalised Tensor Decomposition, i.e., a machine learning approach that can provide forecasts for multiple populations at the same time. Our main conclusions are that single-task and multi-task neural networks provide competitive performance in terms of forecasting errors when compared to the other methodologies considered here. Furthermore, adding a weighting scheme in the training of the neural networks enhances their performance with reference to the mean absolute percentage error.
The remainder of this chapter is organised as follows: Section 3.2 explores existing methods used to forecast cause-of-death mortality rates. Section 3.3 explains the single-task and multi-task neural networks implemented in this study. Section 3.4 describes the data used for the quantitative analysis. Section 3.5 quantitatively compares the out-of-sample performance of existing methodologies with the neural networks we implemented. Section 3.6 addresses an issue that emerged in the previous section by introducing a weighting scheme into the training of the neural networks. Finally, Section 3.7 presents the conclusions and offers a potential extension for this research.

3.2 Existing methodologies

As our main focus is out-of-sample mortality forecasting, we present here the two approaches that are described in the literature as the best ones for this objective, see Zhang et al. (2023). In the following sections of this chapter, we will use these two methodologies as benchmarks against which we compare the neural networks method that we are going to implement.

3.2.1 Lee-Carter model and extensions

Based on the literature (Zhang et al. (2023)), traditional stochastic models such as the Lee-Carter model Lee and Carter (1992), the Renshaw-Haberman model (Renshaw and Haberman (2006)) and the Cairns-Blake-Dowd model (Cairns et al. (2006)), are considered as benchmarks, on which it is difficult to improve, when our goal is forecasting specific cause-of-death mortality rates.
In all these models, the stochastic number of deaths $D_{x,t}^{(g,c)}$ by cause c, in year t, at age x and gender g, follows a binomial distribution

$$D_{x,t}^{(g,c)} \sim Binomial(q_{x,t}^{(g,c)}, N_{x,t}^{(g)}) \tag{3.1}$$

where $N_{x,t}^{(g)}$ is the number of people of gender g and age x alive at the beginning of the calendar year t, and $q_{x,t}^{(g,c)}$ is their one year probability of death for the cause of death c. The probabilities $q_{x,t}^{(g,c)}$ are obtained, on a logit scale, as a linear combination of stochastic factors depend on calendar and/or cohort year, and age-dependent parameters. For the Lee-Carter model, we have:

$$\ln\left(\frac{q_{x,t}^{(g,c)}}{1-q_{x,t}^{(g,c)}}\right) = \alpha_x^{(g,c)} + \beta_x^{(g,c)}\kappa_t^{(g,c)}, \tag{3.2}$$

where $\kappa_t^{(g,c)}$ is the stochastic factor depending by the calendar year, while $\alpha_x^{(g,c)}$ and $\beta_x^{(g,c)}$ are the age-dependent parameters. These quantities are usually estimated via numerical optimization algorithms in order to maximize the corresponding likelihood using observed numbers of deaths. Finally, in order to obtain forecasts for the future probabilities of dying, we fit some time series stochastic process on the stochastic factors and extrapolate them beyond the last year of observation.
As previously mentioned, this approach is reported to have very good performance in terms of forecasting accuracy (Zhang et al. (2023)). Nevertheless, it present the problem that it forecasts mortality rates for different causes of death independently, without taking into account any type of correlation or dependence among the different causes of death.

3.2.2 Tensor decomposition

A tensor is a multi-dimensional array, where the number of dimensions is referred to as the order or degree (d). For example, a tensor is a vector when $d = 1$, a matrix when $d = 2$, and 3-dimensional array when $d = 3$, etc. Tensor decomposition is the process, through a sequence of elementary operations, of breaking down a tensor into simpler, more meaningful parts, facilitating easier analysis and interpretation of complex data, see Kolda and Bader (2009). Tensor decomposition can be seen as an extension of techniques such as Principal Component Analysis and Singular Value Decomposition, see Abdi and Williams (2010), which are used for dimension reduction and feature extraction in two-dimensional data, while tensor decomposition extends these concepts to higher-dimensional data. This extension allows for the capture of more intricate relationships within the data.
Let us assume we have an $N-$dimensional tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$, where d_n is the size of dimension n, representing a set of data. The aim is to find an $N-$dimensional tensor $\mathcal{M} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ that approximates the tensor $\mathcal{X}$. There are several proposals for achieving this goal, including techniques such as CANDECOMP/PARAFAC, Carroll

and Chang (1970), where $\mathcal{M}$ can be written as the sum of r outer products $\mathbf{a}_l^1 \otimes \cdots \otimes \mathbf{a}_l^N$ of rank-one tensors. In matrix notation:

$$\mathcal{M} = \sum_{l=1}^{r} \mathbf{a}_l^1 \otimes \cdots \otimes \mathbf{a}_l^N,$$

where $\mathbf{a}_l^n$, $n = 1, \ldots, N$ is a rank-one tensor (vector) of dimension d_n, for $l = 1, \ldots, r$. The goal of tensor decomposition is to find $\mathcal{M}$ in order to minimize the sum of squared errors

$$||\mathcal{X} - \mathcal{M}||^2$$

where $||.||$ is the norm of a tensor, or other error metrics.
Another proposal is Tucker decomposition, Tucker (1966), where the tensor $\mathcal{X}$ is approximated by a tensor $\mathcal{M}$ which is decomposed into a core tensor and factor matrices:

$$\mathcal{M} = G \times_1 A^{(1)} \times_2 A^{(2)} \times_3 \cdots \times_N A^{(N)},$$

where $G \in \mathbb{R}^{R_1 \times \cdots \times R_N}$ is a core tensor of reduced dimensions; $A^{(n)} \in \mathbb{R}^{d_n \times R_n}$ are the factor matrices for each mode $n = 1, \ldots, N$; $R_n < d_n$ are the ranks (or dimensions) of the core tensor along each mode; and $\times_n$ denotes the mode-n product, which multiplies the core tensor by the matrix along the n-th mode.

In the mortality forecasting environment, the $N-$dimensional tensor is generally containing central mortality rates or probabilities of dying, and the first two dimensions represent age and calendar year, while the remaining can represent the remaining variables such as gender, country, cause of death, socio-economic class, etc.. For example, Russolillo et al. (2011) consider a $3-$dimensional tensor, to which they apply Tucker decomposition, where the third dimension represents different European countries. Once the vector representing the calendar year is calculated, it is forecasted beyond the last year of observation using some time series approach such as ARIMA, linear extrapolation, or smoothing.

Adaptive Penalized Tensor Decompositions

Adaptive Penalized Tensor Decomposition is a particular type of Tensor Decomposition in which it is incorporated a penalization mechanism to adaptively enforce certain properties or constraints on the decomposition, see Madrid-Padilla and Scott (2017). This technique has been used by Zhang et al. (2023) in order to have simultaneous forecasts of mortality rates by cause of death. It belongs to the CANDECOMP/PARAFAC decompositions family. It has been proven to provide good results on a US mortality dataset.
In the following sections of this chapter, we consider both the case of 3-dimensional tensor (year, age, and either cause-of-death or socio-economic class) and 4-dimensional tensor (year, age, cause-of-death, and socio-economic class). The variable gender is treated separately. The number of components r is set equal to 11, following Zhang et al. (2023), and the forecasts are obtained using Generalized Additive Model, i.e. a forecasting method for non-linear but still smooth variables.

3.3 Neural Networks

Our aim is to implement feed-forward neural networks on a mortality dataset for forecasting the mortality rates by cause-of-death c and/or socio-economic class i, where

$c = 1, \ldots, C$, and $i = 1, \ldots, I$. Similarly to what was done in Chapter 2, the neural networks will be of both single-task and multi-task type.

3.3.1 Single-task NNs

The structure of the NN is

- Input layer:
$$\tilde{\mathbf{X}} = (t, x, g, c, i) \in \mathbb{R}^5, \tag{3.3}$$
where t, x, g, c, and i are respectively the calendar year, age, gender, cause-of-death, and socio-economic class.

- Embedding layer:
$$\mathbf{X} = (t, \mathbf{x}, \mathbf{g}, \mathbf{c}, \mathbf{i}) \in \mathbb{R}^{21}, \tag{3.4}$$
where $t \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^5$, $\mathbf{g} \in \mathbb{R}^5$, $\mathbf{c} \in \mathbb{R}^5$, and $\mathbf{i} \in \mathbb{R}^5$ are respectively the calendar year, embedded age vector, embedded gender vector, embedded cause-of-death vector, and embedded socio-economic class.

- Hidden layer 1:
$$\mathbf{Z}^{(1)} = f^{(1)}(\mathbf{c}_1 + \mathbf{B}^{(1)}\mathbf{X}) \in \mathbb{R}^{128}, \tag{3.5}$$
where $\mathbf{c}_1 \in \mathbb{R}^{128}$, $\mathbf{B}^{(1)} \in \mathbb{R}^{128\times 21}$, and $f^{(1)} = \tanh$.

- Hidden layer 2:
$$\mathbf{Z}^{(2)} = f^{(2)}(\mathbf{c}_2 + \mathbf{B}^{(2)}\mathbf{Z}^{(1)}) \in \mathbb{R}^{128}. \tag{3.6}$$
where $\mathbf{c}_2 \in \mathbb{R}^{128}$, $\mathbf{B}^{(1)} \in \mathbb{R}^{128\times 128}$, and $f^{(2)} = \tanh$.

- Output layer:
$$\hat{Y} = Z^{(3)} = f^{(3)}(c_3 + \mathbf{B}^{(3)}\mathbf{Z}^{(2)}) \in \mathbb{R}. \tag{3.7}$$
where $c_3 \in \mathbb{R}$, $\mathbf{B}^{(3)} \in \mathbb{R}^{1\times 128}$, $f^{(3)} = \mathit{Sigmoid}$, and $\hat{Y}$ represents the estimated mortality rate.

3.3.2 Multi-task NNs

Here we implement a multi-task NN with C different tasks, where the task c, for $c = 1, \ldots, C$ consists of forecasting mortality rates for the cause-of-death c.
The structure of the multi-task NN is the following

- Embedding layer:
$$\mathbf{X}_c = (t, \mathbf{x}, \mathbf{g}, \mathbf{c}, \mathbf{i}) \in \mathbb{R}^{21}, \quad c = 1, \ldots, C, \tag{3.8}$$
$$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_C) \in \mathbb{R}^M, \tag{3.9}$$
where $M = 21 \cdot C$.

- Hidden layer 1:
$$\mathbf{Z}^{(1)} = f^{(1)}(\mathbf{c}^{(1)} + \mathbf{B}^{(1)}\mathbf{X}) \in \mathbb{R}^{128}, \tag{3.10}$$
where $\mathbf{c}^{(1)} \in \mathbb{R}^{128}$, $\mathbf{B}^{(1)} \in \mathbb{R}^{128\times M}$, and $f^{(1)} = \tanh$.

- Hidden layer 2:

$$\mathbf{Z}^{(2)} = f^{(2)}(\mathbf{c}^{(2)} + \mathbf{B}^{(2)}\mathbf{Z}^{(1)}) \in \mathbb{R}^{128}, \tag{3.11}$$

where $\mathbf{c}^{(2)} \in \mathbb{R}^{128}$, $\mathbf{B}^{(2)} \in \mathbb{R}^{128\times128}$, and $f^{(2)} = \tanh$.

- Cause-of-death specific layers:

$$\mathbf{Z}_c^{(3)} = f^{(3)}(\mathbf{c}_c^{(3)} + \mathbf{B}_c^{(3)}\mathbf{Z}^{(2)}) \in \mathbb{R}^{32}, \quad c = 1, \ldots, C, \tag{3.12}$$

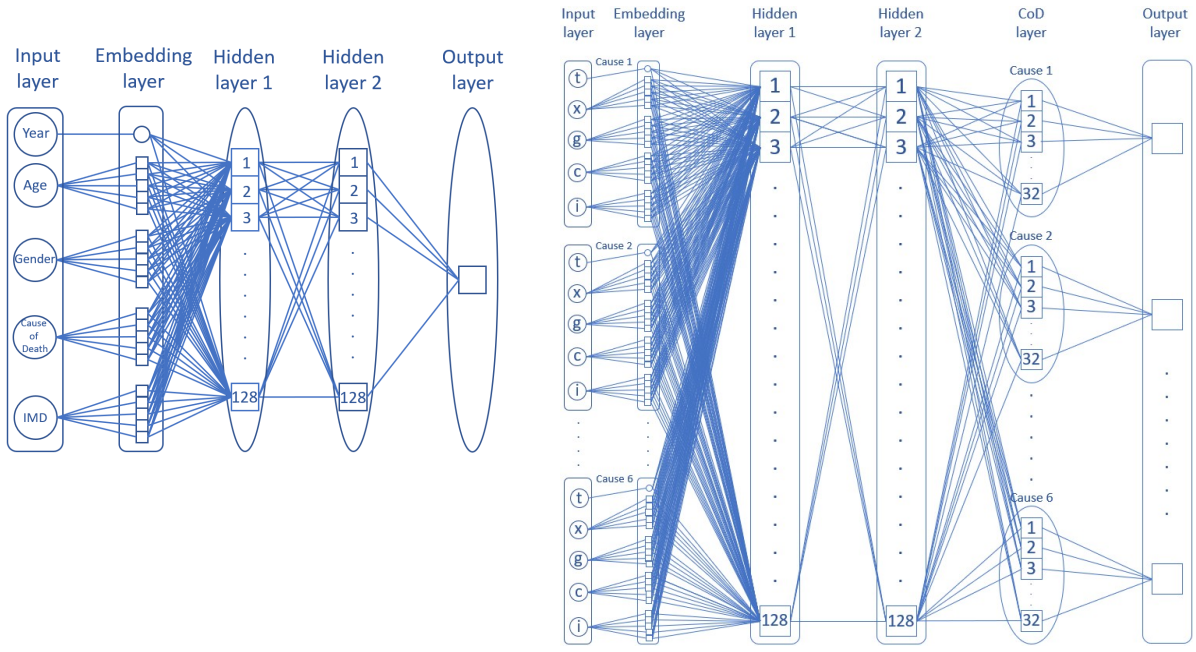
where $\mathbf{c}_c^{(3)} \in \mathbb{R}^{32}$, $\mathbf{B}_c^{(3)} \in \mathbb{R}^{32\times128}$, and $f^{(3)} = \tanh$.

- Output layers:

$$Z_c^{(4)} = f^{(4)}(c_c^{(4)} + \mathbf{B}_c^{(4)}\mathbf{Z}_c^{(3)}) \in \mathbb{R}, \quad c = 1, \ldots, C, , \tag{3.13}$$

where $\mathbf{c}_c^{(4)} \in \mathbb{R}$, $\mathbf{B}_c^{(4)} \in \mathbb{R}^{1\times32}$, and $f^{(4)} = \mathit{Sigmoid}$.

A graphical representation of the two neural networks can be found in Figure 3.3.



(a) Single-task NN. (b) Multi-task NN.

Figure 3.3: Single-task and multi-task neural networks used to forecast mortality rates for 6 specific cause-of-death plus IMD variable.

3.3.3 Multiple Deprivation Index

We describe here a socio-economic index named IMD[1] (Multiple Deprivation Index). This index is a UK government statistical measure used to assess the relative deprivation of neighbourhoods across various domains such as income, employment, health, education,

[1] `https://data.cdrc.ac.uk/dataset/index-multiple-deprivation-imd`

crime, housing, and environment. It combines data from these domains to provide a single deprivation score for each area, helping to identify regions that require targeted policy interventions and resource allocation. In this paper, we consider the quintiles of this index, where the first quintile corresponds to the most deprived neighbourhoods, and the fifth one to the least deprived. In the following sections of this paper, the IMD quintiles will be used as variable representing the socio-economic class.

3.3.4 Considering different input variables

In order to make our analysis more complete, we also implemented restricted versions of the single-task and multi-task NNs described above. Specifically, we considered two cases with respectively 4 and 6 causes of death and without socio-economic class as input variable, a case with socio-economic class as input variable but no cause of death, and finally the complete case with both 6 causes of death and socio-economic class. Table 3.1 resumes the four cases considered. Notice that calendar year, age, and gender are included as input variables in all the cases studied. Finally, see at Figures 3.6, 3.7, and 3.8 for graphical representations of the NNs used in the first three cases.

Table 3.1: Cases considered for neural networks implementation.

Case	No. causes of death	No. socio-economic classes
1	4	-
2	6	-
3	-	5
4	6	5

3.4 Data

The data have been provided by the ONS (Office for National Statistics). The dataset contains the number of individuals alive in England and Wales by age, gender, calendar year, and socio-economic class (IMD) from 2001 to 2020, alongside with the number of deaths by cause for each of these categories. Table 3.2 summarizes the causes of death used in the data set, and the corresponding ICD-10 codes, i.e. alphanumeric codes used by World Health Organization to classify medical conditions and diseases. In the following, we have considered two cases. In the first one, the causes of death are grouped in 4 clusters: Cluster 1 = Neoplasms. Cluster 2 = Diseases of the respiratory system. Cluster 3 = Diabetes, Intentional injuries, Unintentional injuries, and All other causes (all other ICD-10 codes). Cluster 4 = Ischaemic heart disease, Stroke/cerebrovascular, and Other circulatory. In the second one, the causes of death are grouped in 6 clusters: Cluster 1 = Neoplasms. Cluster 2 = Diseases of the respiratory system. Cluster 3 = Diabetes and All other causes (all other ICD-10 codes). Cluster 4 = Ischaemic heart disease and Other circulatory. Cluster 5 = Stroke/cerebrovascular. Cluster 6 = Intentional injuries and Unintentional injuries.

For all the approaches considered, we used as training period: 2001-2015, test period: 2016-2020, and age interval: 25-89. We did not consider the 0-24 age range because we are not interested in childhood mortality, and 90+ age range because of data quality issues.

Table 3.2: Underlying cause of death groups and related codes.

Underlying cause group	ICD-10 codes
01 Neoplasms	C00-D48
02 Diseases of the respiratory system	J00-J99
03 Diabetes	E10-E14
04 Ischemic heart disease	I20-I25
05 Stroke/cerebrovascular	I60-I69
06 Other circulatory	I00-I19, I26-I59, I70-I99
07 Intentional injuries	X60-Y09, Y35-Y36, Y87, Y89, U011
08 Unintentional injuries	V01-V06, V09-V99, W00-W46, W49-W60, W64-W70, W73-W81, W83-W94, W99-X06, X08-X40, X43-X44, X46-X48, X50-X54, X57-X59, Y40-Y66, Y69-Y86, Y88
09 All other causes (all other ICD-10 codes)	All other ICD-10 codes not included in one of the other definitions.
10 Neonatal	All deaths under 28 days.

Finally, in Table 3.3, the hyper-parameters used for the training of the single-task (ST) and multi-task (MT) neural networks are reported. Notice that the training of all the NNs has been repeated ten times in order to balance the randomicity of the training period that allocated random values as initial values of the parameters.

Table 3.3: Hyper-parameters used for the training of the neural networks in all the cases considered.

Case	NN	Loss Function	Optimizer	Batch Size	Validation Set	Epochs	Learning Rate
4 Cause-of-Death	**ST**	MSE	Adam	32	0.05	1500	0.001
	MT	MSE	Adam	32	0.05	1100	0.0005
6 Cause-of-Death	**ST**	MSE	Adam	32	0.05	300	0.001
	MT	MSE	Adam	32	0.05	1200	0.0005
5 IMD	**ST**	MSE	Adam	32	0.05	950	0.001
	MT	MSE	Adam	32	0.05	650	0.0005
6 Cause-of-Death + 5 IMD	**ST**	MSE	Adam	32	0.05	50	0.001
	MT	MSE	Adam	32	0.05	350	0.0005

3.5 Results

In order to compare the out-of-sample performance of the different approaches here considered, we consider several metrics. Indicated with N the size of the test set, and with $\hat{y}_i$ and y_i, $i = 1, \ldots, N$, respectively the observed and predicted values of the dependent variable, the mean absolute error is

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|, \tag{3.14}$$

the mean squared error is

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2, \tag{3.15}$$

and the mean absolute percentage error is

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} 100 \cdot \frac{|\hat{y}_i - y_i|}{y_i}. \tag{3.16}$$

Results by case studied, gender, cause of death, and socio-economic class are reported in the first four columns of Tables 3.4-3.15.[2] Meanwhile, Table 3.16 lists the number of parameters by approach and case studied. In all the tables mentioned, the LC model is referred to as LC, tensor decomposition as TD, single-task NN as ST, and multi-task NN as MT.

For the 4 cause-of-death case, in Tables 3.4 and 3.5 we notice how LC model is the best approach in terms of total MAE and MSE, although that is not the case for all causes of death, followed by multi-task NN, single-task NN, and tensor decomposition. For all the approaches considered, we observe how the magnitude of the errors are higher for male than female, due to having higher observed dying probabilities, and for "Others" cause of death. Focusing on MAPE, see Table 3.6, it appears to be much higher for NNs compared to LC model and Tensor Decomposition, and in particular it is very high for Respiratory Diseases. This could be an indication that NNs do not work well with categories having smaller observed dying probabilities.

Table 3.4: Overall MAE by gender, cause of death and approach. 4 causes of death. Results are multiplied by 10^4.

	LC	**TD**	**ST**	**MT**	**ST w**	**MT w**
Total	4.10	8.82	5.23	4.81	5.93	4.54
Male	4.78	9.41	5.85	5.53	6.54	5.25
Female	3.42	8.22	4.61	4.09	5.33	3.83
Neoplasms	2.31	2.84	2.80	2.39	2.90	2.76
Respiratory Dis.	3.08	3.28	3.70	3.36	4.25	3.13
Heart Dis.	2.72	11.34	3.20	5.33	6.21	4.10
Others	8.30	17.81	11.22	8.16	10.37	8.16

Table 3.5: Overall MSE by gender, cause of death and approach. 4 causes of death. Results are multiplied by 10^6.

	LC	**TD**	**ST**	**MT**	**ST w**	**MT w**
Total	1.77	8.51	3.06	1.99	2.46	2.07
Male	2.59	9.42	3.88	2.71	2.99	2.77
Female	0.95	7.59	2.25	1.27	1.93	1.37
Neoplasms	0.23	0.22	0.14	0.13	0.25	0.23
Respiratory Dis.	0.71	0.75	0.69	0.71	1.28	0.72
Heart Dis.	0.22	5.69	0.31	1.08	1.74	0.61
Others	5.91	27.36	11.11	6.06	6.57	6.71

For the 6 cause-of-death case, see Tables 3.7 and 3.8, we notice that the LC model remains the best one for total MAE while the multi-task NN is the best one for total MSE, likely due to the fact that the training of the NN aims to minimize the in-sample

[2]In the last two columns of Tables 3.4-3.15, the results obtained using single-task NNs with a weighting scheme (ST NN w) and multi-task NNs with a weighting scheme (MT NN w) are reported. The details of these NNs will be discussed in Section 3.6.

Table 3.6: Overall MAPE by gender, cause of death and approach. 4 causes of death.

	LC	TD	ST	MT	ST w	MT w
Total	14.90	15.82	152.35	84.07	24.33	32.20
Male	13.35	13.92	107.96	67.05	23.37	30.81
Female	16.45	17.72	196.74	101.09	25.28	33.58
Neoplasms	8.67	10.61	57.57	46.87	15.88	28.09
Respiratory Dis.	22.25	21.92	428.38	165.06	43.72	42.10
Heart Dis.	13.92	16.91	96.54	99.81	26.43	37.04
Others	14.77	13.85	26.90	24.54	11.27	21.55

MSE. Tensor Decomposition consistently has the poorest performance. When we use the MAPE metric (see Table 3.9) the results show the same behaviour as the previous case (see Table 3.6), with NNs being clearly outperformed by LC models and Tensor Decomposition. Finally, similar to the 4 cause-of-death scenario, we observe that both MAE and MSE are significantly highest for the "Others" cause of death. This is likely due to the impact of the COVID-19 pandemic in 2020, as COVID-related deaths are included in the 'Others' category.

Table 3.7: Overall MAE by gender, cause-of-death and approach. 6 causes of death. Results are multiplied by 10^4.

	LC	TD	ST	MT	ST w	MT w
Total	2.85	5.92	3.25	3.69	3.20	3.31
Male	3.32	6.77	3.67	4.05	3.63	3.89
Female	2.37	5.07	2.82	3.33	2.76	2.73
Neoplasms	2.31	2.76	2.57	3.17	2.36	3.47
Respiratory Dis.	3.08	3.74	2.96	4.05	2.86	3.11
Heart Dis.	2.22	6.74	2.19	3.97	3.00	3.39
Others	8.02	17.75	9.00	8.17	8.43	7.35
Stroke	0.87	3.74	1.74	2.00	1.66	1.84
Injuries	0.57	0.76	1.03	0.78	0.87	0.71

Table 3.8: Overall MSE by gender, cause-of-death and approach. 6 causes of death. Results are multiplied by 10^6.

	LC	TD	ST	MT	ST w	MT w
Total	1.19	5.06	1.63	1.18	1.27	1.07
Male	1.74	6.10	2.20	1.44	1.75	1.39
Female	0.63	4.03	1.06	0.92	0.79	0.75
Neoplasms	0.23	0.22	0.14	0.36	0.24	0.45
Respiratory Dis.	0.71	1.00	0.48	0.95	0.64	0.71
Heart Dis.	0.15	1.70	0.16	0.49	0.45	0.50
Others	5.98	26.65	8.85	5.05	6.13	4.54
Stroke	0.04	0.78	0.12	0.20	0.13	0.21
Injuries	0.01	0.03	0.03	0.02	0.02	0.02

In the case of five socio-economic classes, as shown in Tables 3.10 and 3.11, we observe that the LC model yields the best performance in terms of total MAE, while the single-task and multi-task neural networks perform best based on total MSE. Across both metrics, it is evident that higher socio-economic classes tend to exhibit smaller forecasting errors, likely due to the lower magnitude of the observed probability of dying. Focusing on MAPE (see Table 3.12), we observe that it increases for all approaches, except for the MT NN, as the socio-economic class improves (from the first quintile to the fifth). Moreover,

Table 3.9: Overall MAPE by gender, cause-of-death and approach. 6 causes of death.

	LC	TD	ST	MT	ST w	MT w
Total	16.08	17.48	128.01	82.08	27.92	30.39
Male	15.15	15.96	112.54	58.07	25.95	28.20
Female	17.02	18.99	143.49	106.08	29.88	32.58
Neoplasms	8.67	10.03	30.75	41.31	15.97	27.35
Respiratory Dis.	22.25	21.89	196.11	141.14	41.38	42.09
Heart Dis.	15.37	17.21	72.26	85.42	29.34	34.37
Others	15.07	15.55	24.10	39.65	16.09	20.71
Stroke	21.26	26.84	417.61	165.76	41.00	41.71
Injuries	13.88	13.32	27.24	19.18	23.70	16.12

MAPE is significantly lower for the LC model and Tensor Decomposition compared to the ST and MT NNs. This is due to the fact that the MAPE gives equal importance to all the observations (while MSE and MAE tends to focus more on the ones with the highest observed values). This indicates that neural networks (both single-task and multi-task) produce higher errors in categories, such as younger ages, with low observed mortality rates. This is further supported by the fact that the MAPE is notably higher for the female population than for the male population when using neural networks. Overall, MAPE is significantly lower for all four approaches in this case compared to the 4 and 6 CoD scenarios.
Comparing Tables 3.9 and 3.12, we observe that the total MAPE for both single-task and multi-task NNs is substantially higher when using cause-of-death data compared with socio-economic data. This can be attributed to the greater variability in mortality rates across causes of death than across socio-economic classes. During training, NNs using mean squared error (MSE) as the loss function tend to underestimate categories with lower-scale mortality rates, resulting in poorer performance for these categories and, consequently, higher MAPE values.

Table 3.10: Overall MAE by gender, IMD and approach. 5 IMDs. Results are multiplied by 10^4.

	LC	TD	ST	MT	ST w	MT w
Total	12.06	13.67	12.55	15.72	10.39	13.14
Male	15.22	16.80	15.02	18.19	12.68	15.77
Female	8.90	10.53	10.08	13.24	8.11	10.52
IMD 1	16.78	17.00	19.81	23.90	15.80	19.35
IMD 2	13.43	12.71	14.08	17.30	11.26	14.46
IMD 3	10.79	13.24	10.93	13.94	9.39	11.95
IMD 4	10.27	13.18	9.91	12.82	8.27	10.91
IMD 5	9.04	12.21	8.00	10.63	7.26	9.05

Finally, in the case considering both 6 causes of death and 5 socio-economic classes, the LC model exhibits the best performance for total MAE, followed by the single-task NN (see Table 3.13), while the multi-task NN achieves the best results for total MSE (see Table 3.14). Among all causes of death, 'others' shows the highest errors. Additionally, we observe the same trend as before, where both MAE and MSE decrease as socio-economic class improves. However, when examining MAPE (see Table 3.15), the trend reverses: larger observed probabilities of dying lead to smaller MAPE. For example, MAPE is higher for females than for males and higher for lower socio-economic classes than for higher ones.
Finally, when comparing the magnitude of errors in the 6 causes of death case (Tables

Table 3.11: Overall MSE by gender, IMD and approach. 5 IMDs. Results are multiplied by 10^6.

	LC	TD	ST	MT	ST w	MT w
Total	11.04	12.98	7.31	7.80	6.15	6.95
Male	16.49	17.30	10.43	10.93	8.79	9.87
Female	5.59	8.66	4.19	4.67	3.50	4.04
IMD 1	20.34	21.83	15.87	15.65	13.79	14.35
IMD 2	13.48	13.53	8.69	9.27	6.70	7.90
IMD 3	8.38	11.54	5.35	6.04	4.67	5.42
IMD 4	7.46	9.82	4.07	4.69	3.28	4.14
IMD 5	5.55	8.17	2.58	3.35	2.30	2.96

Table 3.12: Overall MAPE by gender, IMD and approach. 5 IMDs.

	LC	TD	ST	MT	ST w	MT w
Total	9.22	9.06	24.14	31.17	16.75	31.38
Male	9.45	8.91	17.55	24.05	11.65	29.84
Female	8.99	9.22	30.73	38.29	21.85	32.92
IMD 1	8.02	7.63	21.58	33.89	10.09	28.74
IMD 2	8.35	7.56	23.53	33.47	14.42	31.19
IMD 3	8.71	8.62	23.22	30.94	17.05	32.32
IMD 4	9.95	10.14	24.22	29.56	19.37	32.19
IMD 5	11.05	11.38	28.13	27.99	22.83	32.46

3.7-3.8) and the 6 causes of death plus 5 socio-economic classes case (Tables 3.13-3.14), we observe that, in the second case, errors are generally higher. For instance, the overall MAE and MSE increase respectively from 2.85 and 1.19 to 3.73 and 1.65 for the Lee-Carter model, from 3.25 and 1.63 to 3.81 and 1.90 for the single-task NN, and from 3.69 and 1.18 to 4.13 and 1.40 for the multi-task NN. This means that adding information about socio-economic class does not help to predict the mortality rates by cause of death. A possible explanation for this is that when we consider socio-economic class as an input variable, we actually split the population into five parts, leading to smaller datasets. This results in higher volatility in the historical mortality rates time series, making the predictions more difficult to calculate.

Table 3.13: Overall MAE by gender, cause of death, IMD and approach. 6 causes of death & 5 IMDs. Results are multiplied by 10^4.

	LC	TD	ST	MT	ST w	MT w
Total	3.73	5.37	3.81	4.13	3.64	4.23
Male	4.42	6.45	4.35	4.68	4.30	4.74
Female	3.04	4.29	3.26	3.59	2.98	3.73
Neoplasms	3.65	4.21	3.16	4.03	3.14	3.95
Respiratory Dis.	3.60	4.12	3.69	4.32	3.41	5.04
Stroke	1.45	2.68	2.32	1.78	1.64	2.23
Heart Dis.	3.20	4.43	3.06	3.86	2.89	4.41
Injuries	1.19	1.04	1.22	1.32	1.26	1.06
Others	9.30	15.71	9.38	9.48	9.50	8.72
IMD 1	5.25	7.59	5.40	5.53	5.11	5.83
IMD 2	4.08	5.83	4.06	4.37	3.92	4.49
IMD 3	3.44	5.02	3.45	3.88	3.37	4.01
IMD 4	3.22	4.46	3.20	3.60	3.03	3.65
IMD 5	2.67	3.92	2.91	3.28	2.77	3.19

Table 3.14: Overall MSE by gender, cause of death, IMD and approach. 6 causes of death & 5 IMDs. Results are multiplied by 10^6.

	LC	TD	ST	MT	ST w	MT w
Total	1.65	4.40	1.90	1.40	1.88	1.65
Male	2.41	5.59	2.61	1.76	2.74	2.08
Female	0.88	3.21	1.18	1.03	1.03	1.22
Neoplasms	0.59	0.73	0.31	0.39	0.34	0.53
Respiratory Dis.	0.91	1.21	0.72	1.09	0.71	1.88
Stroke	0.11	0.40	0.22	0.14	0.14	0.34
Heart Dis.	0.41	0.89	0.33	0.46	0.30	0.99
Injuries	0.06	0.04	0.04	0.05	0.05	0.03
Others	7.80	23.15	9.76	6.24	9.77	6.10
IMD 1	3.06	9.27	4.16	2.60	4.09	3.22
IMD 2	1.86	5.12	2.16	1.63	2.19	1.83
IMD 3	1.40	3.45	1.33	1.14	1.34	1.35
IMD 4	1.08	2.55	1.02	0.91	1.01	1.04
IMD 5	0.83	1.62	0.80	0.69	0.78	0.79

Table 3.15: Overall MAPE by gender, cause of death, IMD and approach. 6 causes of death & 5 IMDs.

	LC	TD	ST	MT	ST w	MT w
Total	27.25	24.36	129.61	45.95	31.94	37.18
Male	25.90	22.86	111.65	44.19	29.62	34.60
Female	28.60	25.87	147.56	47.71	34.27	39.76
Neoplasms	18.81	17.81	50.92	38.18	21.13	30.53
Respiratory Dis.	30.46	27.43	205.57	53.23	40.11	48.11
Stroke	36.12	33.15	328.33	54.57	41.41	49.35
Heart Dis.	27.06	23.06	111.97	45.94	33.21	38.82
Injuries	30.22	25.12	54.63	38.58	32.42	28.35
Others	20.81	19.62	26.22	45.19	23.37	27.91
IMD 1	23.94	21.66	92.52	41.76	29.53	32.71
IMD 2	25.41	23.14	118.92	44.04	30.95	35.29
IMD 3	25.88	23.64	127.07	46.95	32.56	37.57
IMD 4	29.15	25.44	143.55	48.24	33.09	39.68
IMD 5	31.85	27.94	165.95	48.74	33.57	40.64

Table 3.16: Number of parameters by approach and case considered.

	4 cause-of-Death	6 Cause-of-Death	5 IMD	6 Cause-of-death + 5 IMD
LC	1,160	1,740	1,450	8,700
TD	1,848	1,892	1,870	2,002
ST	20,771	20,781	20,776	21,446
MT	48,292	63,666	55,974	67,656

3.6 Adding a weighting scheme

The training of a neural network typically involves minimizing the sum of errors, defined as the difference between observed and predicted values. In this process, observations with larger values tend to exert a disproportionately greater influence on the network's parameter adjustments. This bias can lead to underestimation for observations with smaller values, such as the probability of dying at younger ages, as demonstrated in Figure 3.4, which shows the mean absolute percentage error across different age groups, approaches, and cases. The same pattern has been observed in Section 3.5, where categories with lower probabilities of dying have much higher MAPE.
This phenomenon occurs because commonly used loss functions, such as Mean Squared Error or Mean Absolute Error, do not account for the relative scale of observations across categories. Higher-value observations thus produce larger absolute errors, which disproportionately influence the training process and lead the network to prioritize minimizing these larger errors. As a result, predictions for categories with smaller values, such as younger age classes, may suffer in accuracy.

To address this issue and ensure that all categories are represented equally, we propose implementing a weighting scheme in the loss function. Specifically, each error $\hat{y}_i - y_i$ will be multiplied by the reciprocal of the corresponding observation value, so the weight will be $\omega_i = \frac{1}{y_i}$. Finally, the loss function will be, considering it as mean squared error:

$$L = \frac{1}{N} \sum_{i=1}^{N} \omega_i (\hat{y}_i - y_i)^2. \tag{3.17}$$

In the following we repeat the training of all the ST and MT neural networks discussed previously, with the same hyper parameters reported in Table 3.3, in all the cases considered in order to see if there is an improvement in neural networks performances in using a weighting scheme in the training.

Results

Mean absolute errors, mean squared errors and mean absolute percentage errors by case studied are reported in the last two columns of Tables 3.4-3.15. Figure 3.4 compares the MAPE by age group of ST and MT neural networks using a weighting scheme with the Lee-Carter model and Tensor Decomposition. Figure 3.5 shows the effect on total MAE, MSE and MAPE of introducing a weighting scheme for the ST and MT NNs for each of the cases considered.
Focusing on Figure 3.4, while comparing the MAPE obtained with and without a weighting scheme, the greatest improvements occur, as expected, for the younger age groups. This is more evident especially for the 4 and 6 causes of deaths cases, and, especially for the single-task NN, in the case where we consider 6 causes of death plus socio-economic class.
As we can observe in Figure 3.5, adding a weighting scheme to the loss function, sensibly decreases the MAPE for both single-task and multi-task NNs in all the four cases studied. The most notable improvement, i.e. decreasing of MAPE, is for single-task NNs though. At the same time, the mean absolute error and the mean square error remain quite similar overall. In particular, for single-task NNs, the mean square error decreases in all cases considered, while the mean absolute error slightly increases for the 4 causes

of death and 5 socio-economic classes cases, and decreases in the other two cases. For multi-task NNs, we have almost no change for the 4 causes of death and 5 socio-economic classes cases, while for 6 causes of death and 6 causes of death plus socio-economic class, we have slightly smaller mean absolute error and mean square error.
Finally, Tables 3.4-3.15 show that the effect of introducing a weighting scheme during neural network training on out-of-sample MAE and MSE is not uniform across categories. The results reveal both improvements and declines in error magnitudes, reflecting a heterogeneous impact.

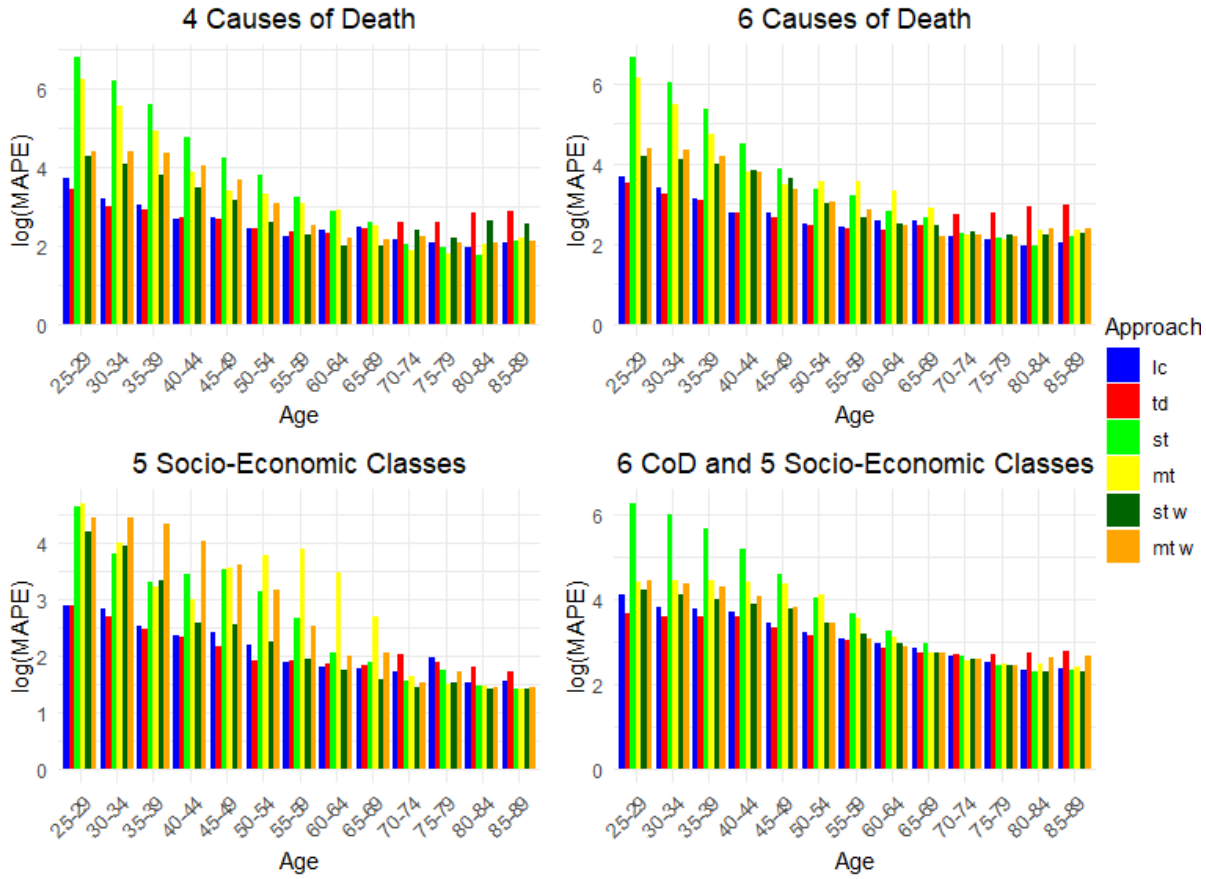


Figure 3.4: log($MAPE$) by age group, approach, and case.

3.7 Conclusion

In this paper, we have implemented single-task and multi-task feed-forward neural networks aimed to forecast mortality rates by cause of death and socio-economic class. To evaluate the out-of-sample performance of these neural networks, an analysis based on an ONS dataset and several metrics has been conducted considering both single-task and multi-task neural networks and also existing methodologies present in the literature such as the Lee-Carter model and Penalized Tensor Decomposition. The most notable result obtained here is that neural networks can achieve similar results in terms of MSE and especially MAE compared with the Lee-Carter model, i.e. the leading model in the literature for out-of-sample performance by cause of death.
Additionally, we notice that neural networks have poorer performance in terms of MAPE

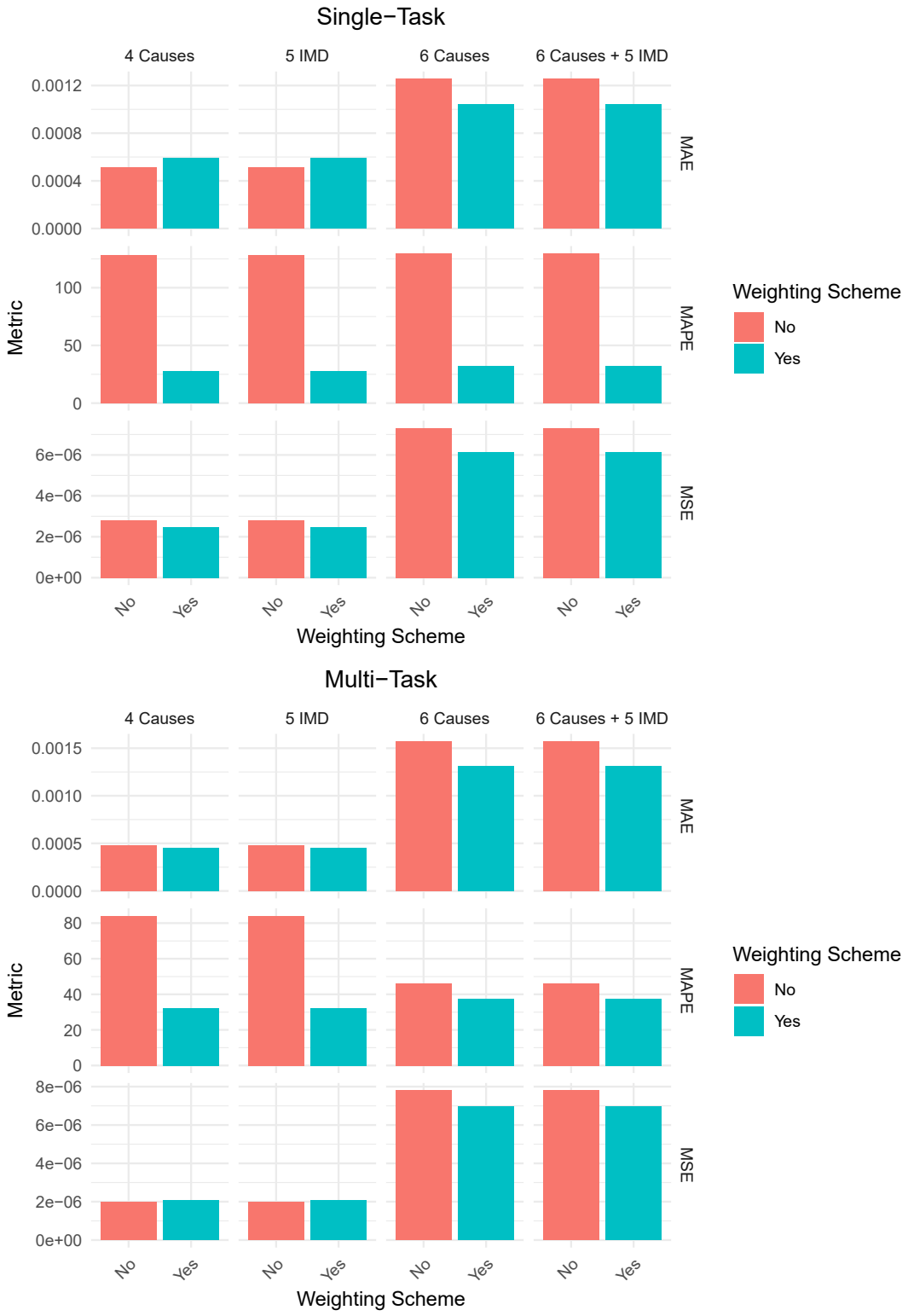
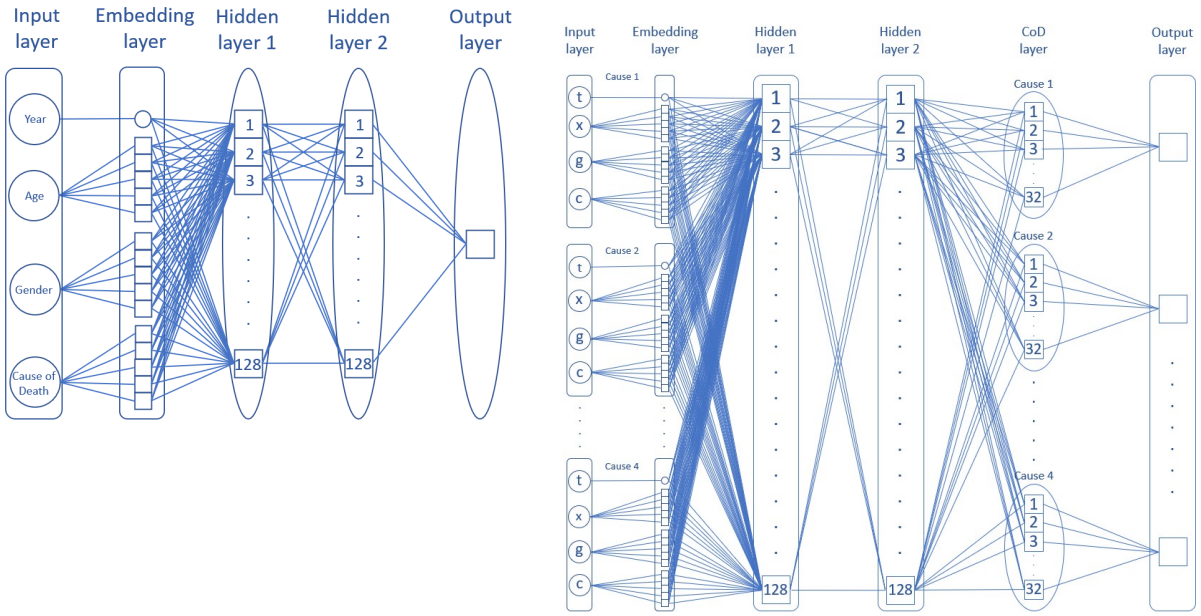


Figure 3.5: Comparison of performance between weighted and non-weighted neural networks across cases and metrics (Total MAE, Total MSE, and Total MAPE).

compared to the Lee-Carter model and Penalized Tensor Decomposition. As this critical aspect is expected to be due to underestimation of categories such as younger ages or less deadly causes of death during the training period due to lower observed probability of dying, we considered a weighting scheme to see if this could help to prevent this issue. We

conclude that this implementation notably improves the results in terms of the MAPE while not changing significantly the results in terms of the MSE and MAE.
As a possible extension of this research, we want to evaluate the possibility of implementing an additional aspect in the loss function which constrains the probabilities of dying by cause-of-death to sum to the total probability of dying, and evaluating if this provides an improvement to the out-of-sample performance.

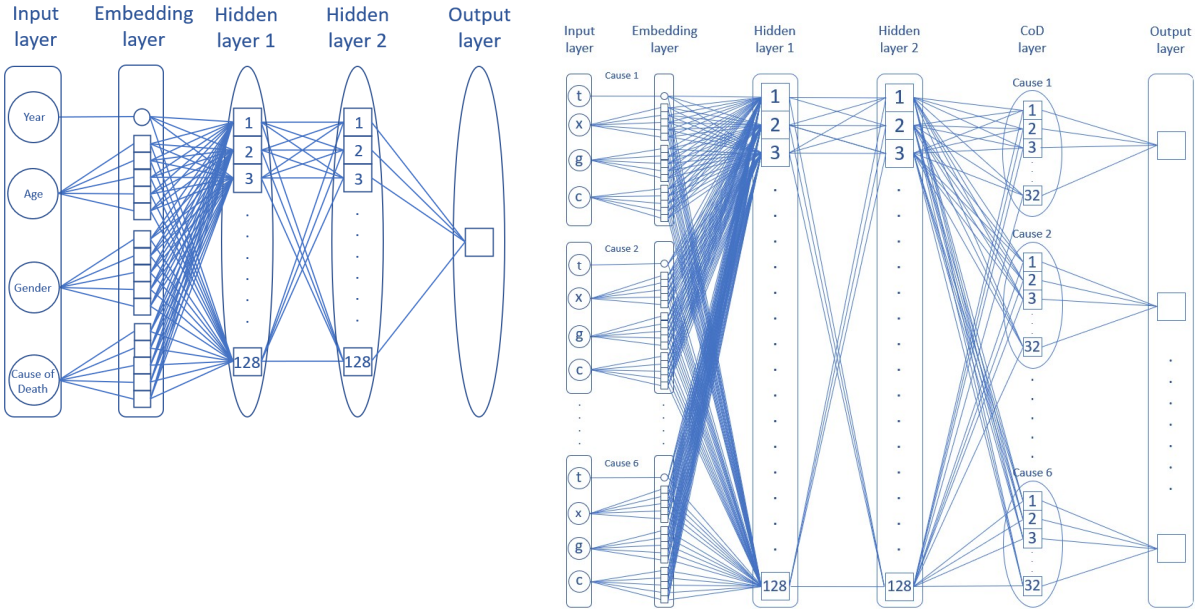
Appendix - Graphical representation of neural networks



(a) Single-task NN.

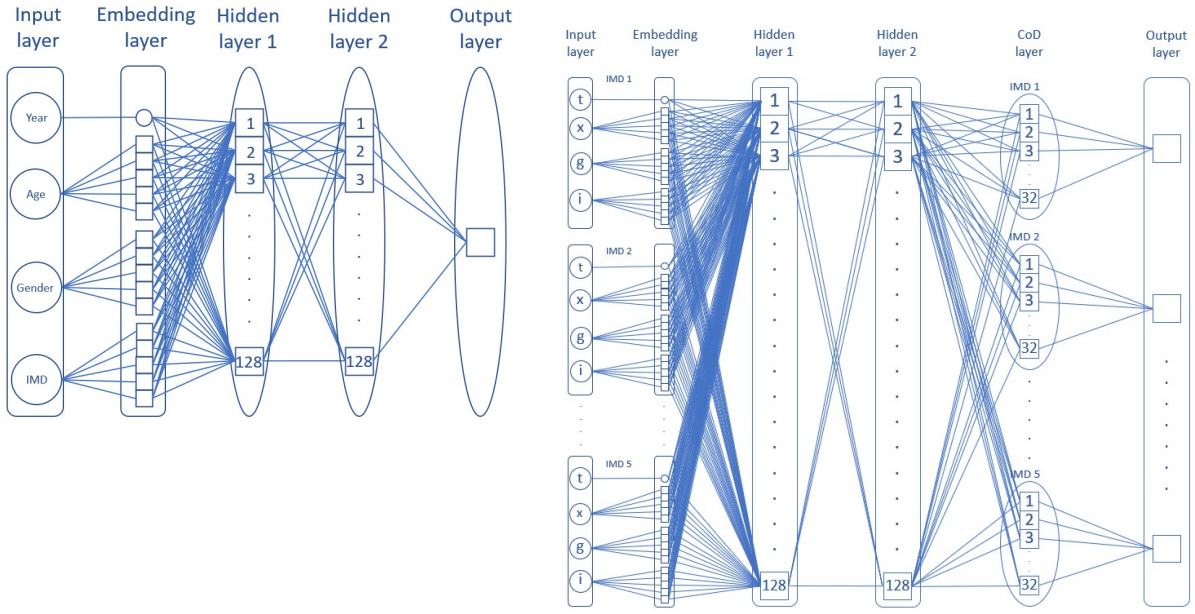
(b) Multi-task NN.

Figure 3.6: Single-task and multi-task NNs used to forecast mortality rates for 4 causes of death.



(a) Single-task NN. (b) Multi-task NN.

Figure 3.7: ST and MT NNs used to forecast mortality rates for 6 causes of death.



(a) Single-task NN. (b) Multi-task NN.

Figure 3.8: ST and MT NNs used to forecast mortality rates for 5 socio-economic classes.

Chapter 4

Conclusion

This thesis has explored different methodological approaches to the modelling and forecasting of mortality, progressing from traditional stochastic frameworks towards data-driven and machine learning–based techniques. Across its three chapters, the research has sought to improve multi-population mortality forecasts through model averaging, multi-task learning, and the incorporation of additional population dimensions such as cause of death and socio-economic class. The results obtained contribute to the growing literature on multi-population mortality forecasting by highlighting both the benefits and the limitations of these diverse approaches.

In the first chapter, the focus was on the comparative performance of two-population stochastic mortality models and several model averaging techniques. Using data from ten countries and various evaluation periods, it was shown that model-averaging approaches generally outperform individual models in terms of both point and interval forecast accuracy. These methods demonstrated enhanced robustness to the choice of metric, country, and time period, suggesting that combining models can effectively mitigate model-specific weaknesses. This chapter therefore provided a solid empirical justification for model averaging as a practical strategy in multi-population mortality forecasting.

Building on this, the second chapter extended the analysis to a deep learning framework by comparing the out-of-sample performance of multi-task neural networks (NNs) in a multi-population context with that of single-task NNs and traditional stochastic models. The findings showed that the performance of multi-task NNs is highly sensitive to the selected age range and training period. While these networks achieved competitive results under certain conditions, their performance tended to decline with broader age ranges due to the underestimation of lower-age mortality. Introducing a weighting scheme significantly improved performance, particularly for life expectancy and standard deviation. The chapter also identified several promising directions for future work, including penalisation schemes to ensure coherence between populations and the use of alternative clustering or machine learning techniques tailored to time series.

The third chapter further expanded the neural network framework by incorporating mortality disaggregation by cause of death and socio-economic class. The analysis demonstrated that both single-task and multi-task NNs can achieve forecasting accuracy comparable to leading models such as Lee–Carter and Penalised Tensor Decomposition in terms of mean and absolute errors. Nonetheless, neural networks exhibited higher mean absolute percentage errors, mainly due to difficulties in learning patterns for less frequent causes of death. The implementation of a weighting scheme mitigated this limitation, improving mean absolute percentage errors without sacrificing overall performance. This

chapter also suggested methodological refinements to ensure internal consistency between cause-specific and total mortality.

Taken together, the three chapters collectively demonstrate a gradual methodological transition—from traditional statistical modelling to the integration of machine learning methods—in the pursuit of more accurate and flexible multi-population mortality forecasts. The results highlight that while traditional models remain competitive and interpretable, combining them with modern computational techniques can yield tangible improvements. Moreover, the application of neural networks reveals the potential of data-driven methods to capture complex dependencies and heterogeneous population structures, provided that appropriate weighting and regularisation strategies are implemented.

Overall, the thesis underscores the importance of combining methodological rigour with computational innovation in demographic forecasting. Future research could focus on hybrid approaches that integrate the interpretability of stochastic models with the flexibility of machine learning, as well as on expanding multi-population and multi-dimensional forecasting frameworks to encompass a broader range of demographic and socio-economic factors.

References

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Alai, D. H., Arnold, S., and Sherris, M. (2015). Modelling cause-of-death mortality and the impact of cause-elimination. *Annals of Actuarial Science*, 9(1):167–186.

Arnold, S. and Sherris, M. (2013). Forecasting mortality trends allowing for cause-of-death mortality dependence. *North American Actuarial Journal*, 17(4):273–282.

Benchimol, A. G., Alonso, P. J., Marín Díazaraque, J. M., and Albarrán Lozano, I. (2016). Model uncertainty approach in mortality projection with model assembling methodologies. *Universidad Carlos III de Madrid. Departamento de Estadística*.

Cairns, A. J., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., and Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, 13(1):1–35.

Carracedo, P., Debón, A., Iftimi, A., and Montes, F. (2018). Detecting spatio-temporal mortality clusters of European countries by sex and age. *International Journal for Equity in Health*, 17(1):1–19.

Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319.

Caselli, G., Vallin, J., and Marsili, M. (2019). How useful are the causes of death when extrapolating mortality trends. An update. *Old and new perspectives on mortality forecasting*, page 237.

Chen, R. Y. and Millossovich, P. (2018). Sex-specific mortality forecasting for UK countries: a coherent approach. *European Actuarial Journal*, 8(1):69–95.

Chen, Y. and Khaliq, A. Q. (2022). Comparative study of mortality rate prediction using data-driven recurrent neural networks and the Lee–Carter model. *Big Data and Cognitive Computing*, 6(4):134.

Clouston, S. A., Rubin, M. S., Phelan, J. C., and Link, B. G. (2016). A social history of disease: contextualizing the rise and fall of social inequalities in cause-specific mortality. *Demography*, 53(5):1631–1656.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Crawshaw, M. (2020). Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.

Danesi, I. L., Haberman, S., and Millossovich, P. (2015). Forecasting mortality in subpopulations using Lee–Carter type models: A comparison. *Insurance: Mathematics and Economics*, 62:151–161.

De Mori, L., Millossovich, P., Zhu, R., and Haberman, S. (2024). Two-population mortality forecasting: An approach based on model averaging. *Risks*, 12(4):60.

Deng, L., Hinton, G., and Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8599–8603. IEEE.

Deprez, P., Shevchenko, P. V., and Wüthrich, M. V. (2017). Machine learning techniques for mortality modeling. *European Actuarial Journal*, 7:337–352.

Dickson, D. C., Hardy, M. R., and Waters, H. R. (2019). *Actuarial mathematics for life contingent risks*. Cambridge University Press.

Djeundje, V. B., Haberman, S., Bajekal, M., and Lu, J. (2022). The slowdown in mortality improvement rates 2011-2017: a multi-country analysis. *European Actuarial Journal*, 12(2):839–878.

Dong, Y., Huang, F., Yu, H., and Haberman, S. (2020). Multi-population mortality forecasting using tensor decomposition. *Scandinavian Actuarial Journal*, 2020(8):754–775.

Dowd, K., Cairns, A. J., Blake, D., Coughlan, G. D., Epstein, D., and Khalaf-Allah, M. (2010). Evaluating the goodness of fit of stochastic mortality models. *Insurance: Mathematics and Economics*, 47(3):255–265.

Dowd, K., Cairns, A. J., Blake, D., Coughlan, G. D., and Khalaf-Allah, M. (2011). A gravity model of mortality rates for two related populations. *North American Actuarial Journal*, 15(2):334–356.

Dubey, S. R., Singh, S. K., and Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, 503:92–108.

Enchev, V., Kleinow, T., and Cairns, A. J. (2017). Multi-population mortality models: fitting, forecasting and comparisons. *Scandinavian Actuarial Journal*, 2017(4):319–342.

Euthum, M., Scherer, M., and Ungolo, F. (2024). A neural network approach for the mortality analysis of multiple populations: a case study on data of the Italian population. *European Actuarial Journal*, 14(2):495–524.

Fletcher, D. (2018). *Model averaging*. Springer Berlin, Heidelberg.

Gaille, S. and Sherris, M. (2011). Modelling Mortality with Common Stochastic Long-Run Trends. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 36:595–621.

Girshick, R. (2015). Fast r-cnn in proceedings of the ieee international conference on computer vision (pp. 1440–1448). *Piscataway, NJ: IEEE.[Google Scholar]*, 2.

Goldberg, Y. (2017). *Neural network methods for natural language processing*. Morgan & Claypool.

Hainaut, D. (2018). A neural-network analyzer for mortality forecast. *ASTIN Bulletin: The Journal of the IAA*, 48(2):481–508.

Hinne, M., Gronau, Q. F., van den Bergh, D., and Wagenmakers, E.-J. (2020). A conceptual introduction to bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, 3(2):200–215.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Jones, W. K., Hahn, R. A., Parrish, R. G., Teutsch, S. M., and Chang, M.-H. (2020). Male mortality trends in the United States, 1900-2010: progress, challenges, and opportunities. *Public Health Reports*, 135(1):150–160.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3):455–500.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American statistical association*, 87(419):659–671.

Levantesi, S. and Pizzorusso, V. (2019). Application of machine learning to mortality modeling and forecasting. *Risks*, 7(1):26.

Li, H. and Lu, Y. (2019). Modeling cause-of-death mortality using hierarchical Archimedean copula. *Scandinavian Actuarial Journal*, 2019(3):247–272.

Li, J. (2013). A Poisson common factor model for projecting mortality and life expectancy jointly for females and males. *Population studies*, 67(1):111–126.

Li, J. S.-H., Zhou, R., and Hardy, M. (2015). A step-by-step guide to building two-population stochastic mortality models. *Insurance: Mathematics and Economics*, 63:121–134.

Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, 42(3):575–594.

Lindholm, M. and Palmborg, L. (2022). Efficient use of data for LSTM mortality forecasting. *European Actuarial Journal*, 12(2):749–778.

Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M. V. (2023). A multi-task network approach for calculating discrimination-free insurance prices. *European Actuarial Journal*, 14(2):1–41.

Lu, J., Wong, W., and Bajekal, M. (2014). Mortality improvement by socio-economic circumstances in England (1982 to 2006). *British Actuarial Journal*, 19(1):1–35.

Mackenbach, J. P., Bopp, M., Deboosere, P., Kovacs, K., Leinsalu, M., Martikainen, P., Menvielle, G., Regidor, E., and De Gelder, R. (2017). Determinants of the magnitude of socioeconomic inequalities in mortality: a study of 17 European countries. *Health & place*, 47:44–53.

Madrid-Padilla, O. H. and Scott, J. (2017). Tensor decomposition with generalized lasso penalties. *Journal of Computational and Graphical Statistics*, 26(3):537–546.

Madrigal, A. M., Matthews, F. E., Patel, D., Gaches, A., and Baxter, S. (2011). What longevity predictors should be allowed for when valuing pension scheme liabilities? *British Actuarial Journal*, 16(1):1–38.

McNown, R. and Rogers, A. (1992). Forecasting cause-specific mortality using time series methods. *International Journal of Forecasting*, 8(3):413–432.

Menet, N., Hersche, M., Karunaratne, G., Benini, L., Sebastian, A., and Rahimi, A. (2023). MIMONets: Multiple-input-multiple-output neural networks exploiting computation in superposition. *Advances in Neural Information Processing Systems*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Nandini, S. and Sanjjushri, V. R. (2023). Estimating countries with similar maternal mortality rate using cluster analysis and pairing countries with identical MMR. *arXiv preprint arXiv:2312.04275*.

Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S., and Perla, F. (2019). A deep learning integrated Lee–Carter model. *Risks*, 7(1):1–16.

Perla, F., Richman, R., Scognamiglio, S., and Wüthrich, M. V. (2021). Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, 2021(7):572–598.

Perla, F., Richman, R., Scognamiglio, S., and Wüthrich, M. V. (2024). Accurate and explainable mortality forecasting with the LocalGLMnet. *Scandinavian Actuarial Journal*, 2024(7):739–761.

Perla, F. and Scognamiglio, S. (2023). Locally-coherent multi-population mortality modelling via neural networks. *Decisions in Economics and Finance*, 46(1):157–176.

Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45(3):393–404.

Pollard, J. H. (1987). Projection of age-specific mortality rates. *Population Bulletin of the United Nations*, (21–22):55–69.

Prince, S. J. (2023). *Understanding deep learning*. MIT press.

Purushotham, M., Valdez, E., and Wu, H. (2011). Global mortality improvement experience and projection techniques. *Society of Actuaries*, 11(1):20–45.

Renshaw, A. E. and Haberman, S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and economics*, 38(3):556–570.

Richman, R. (2022). Mind the gap–safely incorporating deep learning models into the actuarial toolkit. *British Actuarial Journal*, 27.

Richman, R. and Wüthrich, M. V. (2021). A neural network extension of the Lee–Carter model to multiple populations. *Annals of Actuarial Science*, 15(2):346–366.

Russolillo, M., Giordano, G., and Haberman, S. (2011). Extending the Lee–Carter model: a three-way decomposition. *Scandinavian Actuarial Journal*, 2011(2):96–117.

Samuels, J. D. and Sekkel, R. M. (2017). Model confidence sets and forecast combination. *International Journal of Forecasting*, 33(1):48–60.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Schnürch, S. and Korn, R. (2022). Point and interval forecasts of death rates using neural networks. *ASTIN Bulletin: The Journal of the IAA*, 52(1):333–360.

Schoen, R. (2013). *Modeling multigroup populations*. Springer Science & Business Media.

Scitovski, R., Sabo, K., Martínez-Álvarez, F., and Ungar, Š. (2021). *Cluster analysis and applications*. Springer.

Scognamiglio, S. (2022). Calibrating the Lee-Carter and the Poisson Lee-Carter models via neural networks. *ASTIN Bulletin: The Journal of the IAA*, 52(2):519–561.

Shang, H. L. (2012). Point and interval forecasts of age-specific life expectancies: A model averaging approach. *Demographic Research*, 27:593–644.

Shang, H. L., Booth, H., and Hyndman, R. J. (2011). Point and interval forecasts of mortality rates and life expectancy: A comparison of ten principal component methods. *Demographic Research*, 25:173–214.

Shang, H. L., Haberman, S., and Xu, R. (2022). Multi-population modelling and forecasting life-table death counts. *Insurance: Mathematics and Economics*, 106:239–253.

Shkolnikov, V. M., Andreev, E. E., and Begun, A. Z. (2003). Gini coefficient as a life table function: computation from discrete data, decomposition of differences and empirical examples. *Demographic Research*, 8:305–358.

Strozza, C., Bergeron-Boucher, M.-P., Callaway, J., and Drefahl, S. (2024). Forecasting inequalities in survival to retirement age by socioeconomic status in Denmark and Sweden. *European Journal of Population*, 40(1):1–28.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Villegas, A. M., Bajekal, M., Haberman, S., and Zhou, L. (2024). Key drivers of long-term rates of mortality improvements in the United States: Period, cohort, and cause of death analysis, 1959–2016. *North American Actuarial Journal*, 28(1):187–217.

Villegas, A. M. and Haberman, S. (2014). On the modeling and forecasting of socioeconomic mortality differentials: An application to deprivation and mortality in England. *North American Actuarial Journal*, 18(1):168–193.

Villegas, A. M., Haberman, S., Kaishev, V. K., and Millossovich, P. (2017). A comparative study of two-population models for the assessment of basis risk in longevity hedges. *ASTIN Bulletin: The Journal of the IAA*, 47(3):631–679.

Villegas, A. M., Kaishev, V. K., and Millossovich, P. (2018). StMoMo: An R package for stochastic mortality modeling. *Journal of Statistical Software*, 84(3):1–38.

Wang, C.-W., Zhang, J., and Zhu, W. (2021). Neighbouring prediction for mortality. *ASTIN Bulletin: The Journal of the IAA*, 51(3):689–718.

Willets, R., Gallop, A., Leandro, P., Lu, J., Macdonald, A. S., Miller, K., Richards, S., Robjohns, N., Ryan, J., and Waters, H. R. (2004). Longevity in the 21st century. *British Actuarial Journal*, 10(4):685–832.

Yang, B., Li, J., and Balasooriya, U. (2016). Cohort extensions of the Poisson common factor model for modelling both genders jointly. *Scandinavian Actuarial Journal*, 2016(2):93–112.

Zhang, X., Huang, F., Hui, F. K., and Haberman, S. (2023). Cause-of-death mortality forecasting using adaptive penalized tensor decompositions. *Insurance: Mathematics and Economics*, 111:193–213.

Zhang, Y. and Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609.